HIT OR MISS? TEST TAKING BEHAVIOR IN MULTIPLE CHOICE EXAMS

Author(s): Pelin Akyol, James Key and Kala Krishna

Source: *Annals of Economics and Statistics*, September 2022, No. 147 (September 2022), pp. 3-50

Published by: GENES on behalf of ADRES

Stable URL: https://www.jstor.org/stable/10.2307/48684785

Annals of Economics and Statistics, Number 147, September 2022

HIT OR MISS? TEST TAKING BEHAVIOR IN MULTIPLE CHOICE EXAMS

PELIN AKYOL[a], JAMES KEY[b] AND KALA KRISHNA[c]

This paper is the first to structurally model how a test taker answers questions in a multiple choice exam. We allow for the possibility of a penalty for a wrong answer which makes risk averse examinees more likely to skip questions. Despite the lack of item response data, we can estimate the model by using the insight that skipping behavior, together with penalties for wrong answers, makes certain scores much more likely than others. Using data from the Turkish University Entrance Exam, we estimate the model and find that candidates' attitudes towards risk differ according to their gender and ability with females and those with high ability being significantly more risk-averse. However, the impact of differences in risk aversion on scores is small. As a result, a higher guessing penalty increases the precision of the exam, and does so with a minimal impact on gender bias.

JEL Codes: I21, J24, D61, C11.
Keywords: Multiple-Choice Exams, Guessing Penalty, Risk Aversion.

1. INTRODUCTION

Multiple-choice test structures are commonly used to evaluate the knowledge of candidates in a wide variety of situations. They are widely used in practice being seen as objective, fair[1] and low cost, especially when large numbers of candidates are involved (Frederiksen (1984) and Becker and Johnston (1999)). University entrance exams in several countries, including Turkey, Greece, Japan, and China, are multiple-choice exams. In the US, the Scholastic Aptitude Tests (SATs) and Graduate Record Exams (GREs) that are taken before applying to undergraduate and graduate schools are also mostly of this form. Such exams are also widely used to measure the effectiveness of schools and teachers, to enter the civil service, and to allocate open positions.[2] Furthermore, scores in such exams are likely to be important determinants of future wages and occupations (Ebenstein, Lavy, and Roth (2016)). Its main advantages are that it allows a broader evaluation of the candidate's knowledge in a short time, it is easy to grade, which matters more when large numbers of candidates are involved, and there is no subjective effect of the grader in the evaluation. Because of these properties, it is preferred in both high and low stake

We would like to thank the Editor, Arnaud Maurel, and two anonymous referees for their valuable comments and suggestions. We also thank Paul Grieco, Sung Jae Jun, Stephen Yeaple and Mark Roberts for their helpful comments on an earlier draft and Lewis Mclean for able research assistance. We would also like to thank seminar and conference participants at the WEAI 11th International Conference, 11th World Congress of the Econometric Society, 30th Congress of the European Economic Association, Conference on the Economics of Health, Education, and Worker Productivity, Massey University Albany, Otago University, Victoria University Wellington, Monash University, Hacettepe University, Sabanci University and Bilkent University.

[a]Bilkent University, Turkey. pelina@bilkent.edu.tr
[b]Frontier Economics, Australia. key.james@gmail.com
[c]Penn State University, CES-IFO and NBER, United States. kmk4@psu.edu

[1]A fair exam is one where the only relevant candidate characteristic is the candidate's knowledge of the material.

[2]For example, in Turkey, public sector jobs are allocated according to the score obtained in a central multiple-choicel exam, called KPSS.

exams in many countries.

A disadvantage of such exams is that candidates may attempt to guess the answer without having any knowledge of the answer[3](see Budescu and Bar-Hillel (1993) and Kubinger, Holocher-Ertl, Reif, Hohensinn, and Frebort (2010)). In other exam types, such as short answer based exams, such uneducated responses are unlikely to reap any benefit. As a response to this problem, test designers may apply negative marking for wrong answers (guessing penalty). Grading methods in multiple-choice tests may be designed in such a way that the expected score from randomly guessing a question is equal to the expected score from skipping the question. This grading method would prevent guessing by risk averse candidates that possess no knowledge as to which answer is correct. However, if they have partial knowledge about the question, the candidate's decision to guess/attempt or skip/omit the question will not only depend on their knowledge, but also on their degree of risk aversion and confidence level.[4] This problem may undermine the validity and the fairness of test scores as different groups may differ in their guessing behavior. If women are less likely to guess, they will have lower expected scores despite having the same ability, which would reduce the efficacy and fairness of the testing mechanism. Guessing would also bias the estimates obtained by item response theory models (IRT) such as the Rasch model, which is the dominant one in the literature, see for example Pekkarinen (2015).[5] The Rasch model, using what is termed item response theory, boils down to predicting the probability of a correct answer using a logit setup, with individual and question fixed effects. The individual fixed effect is thought of as ability, and the question fixed effect as the difficulty of the question. It is used in evaluating data by the OECD for exams that compare performance across countries and over time such as PISA and TIMSS. The PISA manual (OECD, 2009) outlines the many variants of this model that are used by practitioners.

Especially when there is negative marking, so that some guessing is expected, IRT models would give biased results. In this case, the candidate's decision would depend on his unobserved characteristics (risk aversion or confidence) which will violate the unidimensionality[6] assumption of IRT (Ahmadi and Thompson (2012)). Therefore, we argue that the standard approaches to examining the performance of candidates taking multiple-choice exams are inadequate in such settings, and often misleading, as they do not take into account skipping/omission behavior properly. We provide a model that takes into account omission due to risk aversion. By allowing for skipping, and relating this to risk aversion we use all the information in the data, in contrast to the standard Rasch model. By ignoring information on skipping, the Rasch model gives biased estimates of a candidate's ability. To understand why this bias occurs, consider, for example, a setting where there is negative marking and all questions are extremely difficult so that candidates have little idea about the right answer, and all candidates have the same ability, though some are risk averse (and so are more likely to skip a question when they are unsure about it)

[3]For example, with no knowledge of the subject and four options on a question, a candidate would, on average, get 25% correct.

[4]A possible change in the exam grading method is removing penalties for wrong answers. This leads all candidates to answer all questions which would increase the noise associated with the score (see Bereby-Meyer, Meyer, and Flascher (2002) and Kubinger, Holocher-Ertl, Reif, Hohensinn, and Frebort (2010)).

[5] Rasch (1993) has over 12,000 citations in Google Scholar.

[6]Unidimensionality refers to the existence of a single trait or construct underlying a set of measures (Gerbing and Anderson (1988)).

while others are not. Say the risk averse group answers 20 of 80 questions getting 10 right, while the risk neutral one answers 40 of 80 questions getting 15 right. In this case, the Rasch model would estimate the probability of answering correctly for the risk averse group as 1/8 and that for the risk neutral group answering 40 questions and getting 15 right as 3/16. However, the difference in the two would be due to differences in risk aversion rather than ability.[7] Such differences in risk aversion are likely to exist: candidates close to an acceptance cutoff for a highly desirable school may well be very risk averse, and it has been argued, see for example Eckel and Grossman (2008a), Charness and Gneezy (2012) and Croson and Gneezy (2009), that females are more risk averse than males. In a multiple choice exam setting, Baldiga (2014) and Funk and Perrone (2016) show that individuals behave in a risk-averse manner, and females are risk averse relative to males. To disentangle ability and risk aversion, and obtain unbiased estimates of both risk aversion and ability, we need to specify a complete setting, one that includes the choice of skipping the question as done here. It is worth noting that despite the interest in such exams in Psychology, Education, and Economics literature, there is little formal modeling and estimation based on models of individual behavior.[8] Thus, improving on existing methods is vital for understanding what lies behind performance differences of different groups and for policy making.

In this paper, we specify and estimate what we believe is the first structural model of candidates' exam taking behavior and use it to understand performance differences by gender and ability in the Turkish university entrance exam. We use administrative data from the 2002 Turkish University Entrance Exam (ÖSS) in our work. The ÖSS is a highly competitive, centralized examination that is held once a year. It is selective as only about a third of the exam candidates are placed at all, and admission to top programs is extremely competitive. College admission depends on the score obtained in the ÖSS, and the high school GPA[9], with at least 75% of the weight being given to the ÖSS score. There are forty-five questions in each section of the exam, and each question has five possible answers; for each correct answer the candidate obtains one point, and for each wrong answer a penalty of 0.25 points is applied, while no points are awarded/deducted for skipping a question.[10] Students expend significant time and effort to prepare for this exam and have a good understanding of how the system works.

Our model based approach to analyzing exam performance data allows us to estimate student ability and guessing behavior (which captures risk aversion/confidence) when item response data is available. This model is presented and estimated (on a small sample of data on mock exams where we do have item response data) in Appendix A.3 and the estimates obtained are compared to those of other approaches.

[7]The reliance on the Rasch model is part of the reason why the OECD uses certain shortcuts in analyzing its PISA exams (namely treating skipped items as if they were not in the exam) though this does not solve the problem. As there is no structural model that captures skipping behavior (which is endemic in PISA despite the absence of negative marking) they are at a loss as to how to deal with skipped questions.

[8]Other issues relating to the format of questions, such as whether multiple choice questions discriminate by group, are beyond the scope of this paper. Griselda (2022) uses PISA data to show that the gender gap in performance increases with the fraction of multiple choice questions relative to open-ended ones.

[9]The high school GPA is normalized at the school-year level using school level exam scores to make GPAs comparable across schools in each year. This also removes the incentive to inflate high school grades.

[10]Thus, the expected score from a random guess is over the five possible answers is zero.

However, item response data is not available in our administrative data.[11] Notwithstanding, we are able to extend our approach to make it apply to our data under certain conditions (negative marking present, most students attempting all questions and uniform question difficulty) as explained in Section 4 below. As there is negative marking in the OSS exam, the first condition is naturally satisfied. We choose the track (Social Science) and components (namely the Social Science and Turkish parts of the exam) so that most students answer most questions in the exams studied. Answering all questions in the presence of negative marking creates spikes in the distribution of scores which we use to estimate the key parameters.[12] Thus, the additional assumption made in our application is only that question difficulty is constant. This assumption does result in a bias in the estimate of the level of risk aversion as shown in simulations in Table A.7, but for our setting, we show this bias is small. It is important in practice to both researchers and practitioners (like makers of educational policy who tend not to have item response data) that our approach can be applied fruitfully to the no item response data setting.[13]

It is worth noting that even with item response data, we cannot distinguish between greater risk aversion and lower confidence since both would make students less likely to guess, which is what pins down a key parameter, the guessing cutoff, c, in our model. Papers that distinguish between them typically have surveys which elicit confidence levels and risk preferences. We would need a survey designed to elicit at least one of the two. See, for example, Baldiga (2014) and Iriberri and Rey-Biel (2021). Our estimate of the guessing cutoff, c, which is increasing in risk aversion, is termed the "risk aversion cutoff" from here on. It includes both true risk aversion and under-confidence. Since c is what drives behavior, we do not need to distinguish between them for our purposes and our existing counterfactual results would not be affected even if we could estimate them separately. Some papers do differentiate between risk aversion and over-under confidence. Most of these are not in a multiple choice exam setting. Such papers include Buser, Niederle, and Oosterbeek (2014), Kamas and Preston (2012) and van Veldhuizen (2016). In a multiple choice exam setting we are aware of three papers. Of these, two papers, Baldiga (2014) and Iriberri and Rey-Biel (2021), find that risk aversion seems to do a lot more than confidence level.[14] A third paper, Karle, Engelmann, and Peitz (2022), looks at loss aversion rather than risk aversion[15]. Their experiments suggest that both attitudes toward risk as well as the degree of confidence matter.

[11]Item response data is not usually available. PISA and TIMSS data do have item responses and are publicly available. However, they are low stakes exams for students so that students may not take the exams seriously.

[12]Note that we are only able to estimate the structural parameters for students in each predicted score interval (not each agent). We could estimate the key parameters student by student if we had item response data.

[13]Our model could also be extended to incorporate a cost of effort (that could be individual specific) that rises with time spent on the exam. This can help explain the pattern of skipped questions observed in the data even in the absence of negative marking as is the case with PISA data.

[14]Baldiga (2014) finds that there is no difference between men and women in terms of their confidence and risk aversion explains about half of the gap in guessing. Iriberri and Rey-Biel (2021) find that one standard deviation increase in the overconfidence leads to only 0.060143 (0.0137*4.39) standard deviation less omission, however, one standard deviation increases in risk aversion increase omission by 0.7575 (0.202*3.75) standard deviation, a much larger number.

[15]Loss aversion allows for asymmetric risk aversion: namely one may be more risk averse when considering losses than gains.

We focus on whether gender affects risk aversion[16] and the consequences of these differences as well as their implications for policy. Females do seem to guess less at all score levels than males. This is in line with the literature[17] that suggests that females' performance in college and the job market is negatively impacted by their less assertive and more risk averse behavior, see Belzil and Leonardi (2013), Bursztyn, Fujiwara, and Pallais (2017) and Manian and Sheth (2021). Students with low expected scores also tend to guess more. This makes sense as there is a cutoff score to qualify for possible placement, and most likely, a jump up in utility upon becoming eligible for placement. If women are more risk averse than men, there may also be a tradeoff between precision and fairness. Negative marking reduces guessing, thereby increasing accuracy. However, it reduces the expected score of the more risk averse, discriminating against them. Our structural approach allows us to understand the performance of alternative exam designs in terms of precision and fairness through our counterfactual analysis. A central takeaway from our counterfactuals is that the tradeoff between risk aversion and precision is small, at least in our setting. Consequently, we find penalties in multiple-choice exams increase test score precision in terms of measuring ability, without relevantly hurting students with higher risk aversion.

Our estimates of differences in risk aversion by gender contrasts with what Baldiga (2014) finds in a lab experiment setting. What might drive this difference in results? In one of the experiments in Baldiga (2014), the test is not framed, in the other, the test is framed as an SAT exam (high stakes) which is closer in stakes to the Turkish setting we study. She finds that women skip more questions than men in both setups, but this difference is smaller when the exam is framed as an SAT exam. Our results are consistent with what Funk and Perrone (2016) find in a field experiment (using data from a micro exam where the stakes are higher) that differences in risk aversion matter little. This suggests that the difference in the context explains a good deal of the differences in results.[18]

Our results show that candidates' attitudes towards guessing do differ slightly by gender and expected score. We then explore the role of risk aversion on the gender gap in performance. Since females in general tend to guess less often than they should if they were maximizing their expected score, they tend to have lower scores, and less variance in their scores, than otherwise similar males. This tends to make them under-represented at both the top and the bottom end of the score distribution. The consequences of under representation at the top could be particularly relevant when university entrance is very selective. However, we find that risk aversion does not have a significant impact on the gender gap for students of high ability because such students tend to rarely skip questions and it is through skipping behavior that differences in risk aversion are manifested. We also investigate the impact of alternative designs on the precision of the exam. Our results show that increasing the penalty for incorrect answers or increasing the number of questions in the exam improves the ability of exams to sort candidates in terms of their ability/knowledge of the material[19]. More precisely, increasing the number of questions

[16]While estimates of risk aversion may be biased under our approach, estimated differences by gender should not be.

[17]See, for example, Croson and Gneezy (2009), Baldiga (2014) and Funk and Perrone (2016).

[18]In the Turkish setting, not only are stakes are very high, but students are coached in how and when to guess to maximize their score which would tend to make both genders act as if they were less risk averse.

[19]In a recent study, Direr (2020) shows that increasing the number of questions in a multiple choice test

from 45 to 70 in an exam with a guessing penalty of 0.25 would sort candidates similarly to when there are 45 questions with guessing penalty of 1. Negative marking is shown to not only increase precision, but to do so with a minimal impact on gender bias.

1.1. *Related Literature*

The psychology and education literature has long been interested in developing test designs that generate fair results. Baker, Barton, Darling-Hammond, Haertel, Ladd, Linn, Ravitch, Rothstein, Shavelson, and Shepard (2010), for example criticize the use of test results of students to evaluate the value-added of teachers and schools partly because of the measurement error generated by random guessing. Risk attitudes of candidates are recognized to be an important factor in the decision to attempt a question whenever there is uncertainty associated with the outcome (see, Espinosa and Gardeazabal (2013)). In the literature, females are shown to be more risk averse than males in many settings (see Eckel and Grossman (2008b), Agnew, Anderson, Gerlach, and Szykman (2008) and Charness and Gneezy (2012)). To test the hypothesis that females students skip more questions than males since they are more risk averse, Ben-Shakhar and Sinai (1991) investigate test taking strategies of students in Hadassah and PET tests in Israel and find that females do, in fact, tend to skip more questions.

The empirical literature presents mixed evidence on the effects of risk aversion on the exam performance of the candidates. In a field experiment, Funk and Perrone (2016) find that although females are more risk averse relative to males, the differences in risk aversion does not have a significant effect on the differences in exam scores. On the other hand, Baldiga (2014) shows in an experimental setting that females are more risk averse than males and skip more questions, and conditional on students' knowledge of the test material, those who skip more questions tend to perform worse suggesting that such exams will be biased against groups who skip questions rather than guess. Iriberri and Rey-Biel (2021) show that female participants skip significantly more questions than their male counterparts when there is a reward for omitted questions, and this leads to a decrease in the score of females. Similarly, Coffman and Klinowski (2020) show that removal of penalties for wrong answers on the national college entry examination in Chile decreased the gender gap especially among high-achievers. The differences in the characteristics of exams in terms of difficulty and stakes, or the specific cultural/country setting might be the source of these mixed findings (see Riener and Wagner (2017) and Saygin and Atwater (2021)).

Burgos (2004) investigates score correction methods that reward partial knowledge by using prospect theory.[20] He derives a fair rule which is also strategically neutral so that an agent with partial knowledge will answer, while one without any knowledge will not. Similarly, Bernardo (1998) analyzes the decision problem of students in a multiple-choice exam to derive a "proper scoring rule", i.e., one that truthfully elicits the probability of each answer being correct.[21] Espinosa and Gardeazabal (2010) models students' optimal

is an effective way to enhance score efficiency.

[20]Prospect theory describes the way people choose between alternatives when the probabilities associated with them are known taking a behavioral approach such as loss aversion.

[21]Proper scoring rules have been developed to reward partial knowledge where students report the subjective probability of each choice being correct rather than choose one answer so that more information is revealed. There are different types of proper scoring rules, quadratic, spherical, and logarithmic (Bickel

behavior in a multiple-choice exam and derives the optimal penalty that maximizes the validity of the test, i.e., maximizes the correlation between students' knowledge and the test score by simulating their model under distributional assumptions on students' ability, difficulty of questions and risk aversion. Using simulations, the paper argues that the optimal penalty is relatively high. Even though the penalty discriminates against risk averse students, this effect seems to be small compared with the measurement error that it prevents, especially for high ability students which is consistent with our results.

None of these works attempt to estimate ability and risk aversion of agents or to test the implications of their models empirically as we do. We use a simple Bayesian setting where candidates of better ability get a more informative signal about the correct answer. This feature, together with risk aversion, allows us to use the skipping behavior of candidates, as well as their accuracy, to estimate ability and risk aversion. The data we use in this study does not include information on the question-by-question responses. As a result, we cannot directly look at the probability of skipping and getting a correct answer. Despite this, we are able to use information on the distribution of scores in the presence of negative marking to infer skipping tendencies and ability distributions as well as risk aversion, while allowing them to differ across groups. Thus, one of our contributions is to provide a way to estimate structural parameters of the model (with a few additional assumptions) with no data on question-by-question responses. In addition, having a structural model lets us do counterfactual exercises.

We proceed as follows. In the next section, we present an overview of the data and testing environment. The particular patterns seen in the multiple choice tests are discussed in more detail in Section 3. In Section 4, the model is presented. Section 5 details the estimation strategy with the results in Section 6. Section 7 contains counterfactual experiments and Section 8 concludes. Additional tables and figures are presented in Appendix A.1. While the main body of the paper focuses on students in the Social Science track, Appendix A.2 examines those in the Language and Turkish-Math tracks. Appendix A.3 presents the extended model that can be used with item-by-item response data. In Appendix A.4, we present additional simulation results with varying question difficulties.

2. BACKGROUND AND DATA

In this study, we use data from the Turkish university entrance exam. Our main source of data is administrative data from the Student Selection and Placement Center (ÖSYM) and the high schools on a random sample of roughly 10% of the 2002 university entrance exam candidates.[22] The university entrance exam is a multiple-choice exam with four test sections: Turkish, Social Science, Math, and Science[23]. Students are given 180 minutes for 180 questions and can choose how to allocate their time between different test sections; all four sections are taken at the same time. Each section of the test has 45 questions, and

(2010)). The comparisons and the details of these methods are beyond the scope of this paper. In practice, the application of these methods is problematic, especially in large scale exams. Its complexity means that its rules may not be internalized by all students which could create another source of inequality.

[22]The advantage of using the Turkish data is that it is on a real world high stakes setting and there is a lot of data relative to most experiments. The disadvantage is that we do not have item response data, only scores at the student level in each exam.

[23]There is also a separate multiple-choice language exam which is held one week after the main exam. This exam is taken by students who aim to get into college programs such as English literature.

each question has five possible answers. Students get one point for each correct answer, and they lose 0.25 points for each wrong answer. If they skip the question, they receive 0 points. The university entrance exam is a paper-based exam; all students receive the same questions, and they do not receive any feedback on whether their answer is correct or not during the exam. Our data includes students' raw test scores in each test, weighted test scores, high school, high school track, high school GPA, gender, and the number of previous attempts. The second source of data is the 2002 university entrance exam candidate survey. This survey is filled by all students while they are making their application for this exam. This data set has information on students' family income, education level, and time spent and expenditure on preparation. We have around 40,000 students from each high school track, Social Science, Turkish-Math, Science, and Language. Students choose one of the Science, Turkish-Math, Social Science, or Language tracks at the beginning of high school. Students' university entrance exam scores (ÖSS score by track) are calculated as a weighted average of raw scores in each test.

In this paper, we focus on first time taker Social Science track students.[24] The track score in Social Science gives the highest weight (1.8) to the Turkish and Social Science sections of the tests, while Math and Science have a relatively low weight (0.4).[25] As students in the Social Science Track tend to find Math and Science difficult, they tend to complete few questions in these sections of the exam. This is evident from the fact that 58% have a score of less than or equal to zero and 90% have a score less than 3.5. For questions in the Turkish and Social Science sections, students are likely to have partial knowledge, choosing the best answer. Moreover, a substantial fraction of students seem to attempt all questions and this results in spikes at certain scores which helps us to pin down risk aversion as explained in more detail below. In contrast, many math and science questions involve solving a problem so that students are either quite certain of the answer or have failed to solve it (thereby having no information regarding which answer is correct) so that the question is skipped independent of risk aversion. This makes identification of risk aversion impossible with the aggregate data. Therefore, our focus will be on Social Science track students and the Social Science and Turkish sections of their test.

As a robustness check, we also estimate the model with data from the Language track and from the Turkish-Math track. We only use the Language test in the former and the Turkish test in the latter. The results from these analysis are presented in Online Appendix A.2.

College admission is based entirely on an annual, nationwide, central university entrance exam. Most high school seniors take the exam and there is no restriction on retaking.[26] However, the score obtained in a year can be used only in that year. As retaking decision is endogenous, including repeat takers into our analysis will not be informative without controlling for the selection problem, therefore we focus on first time taker students.

[24]We check for robustness by repeating our analysis for second time takers. This can be found in Appendix A.1 Table A.5.

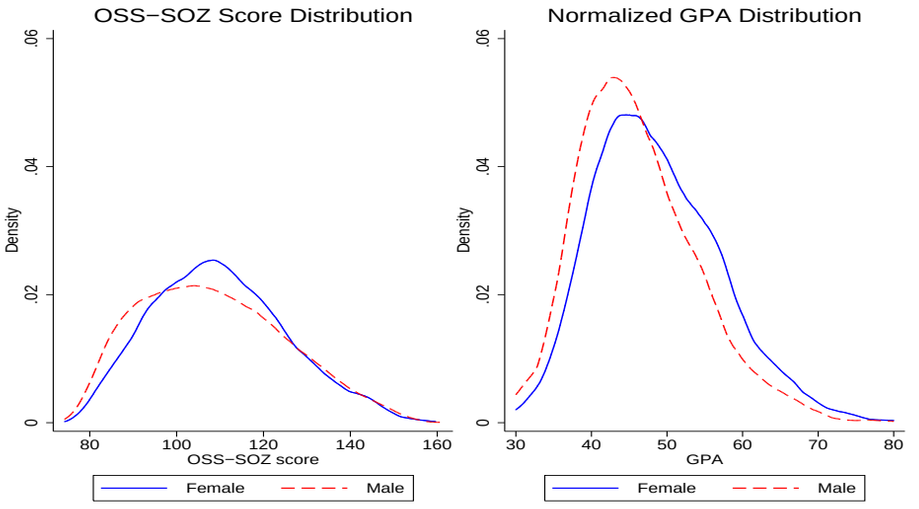
[25]In the calculation of ÖSS scores, firstly raw scores in each track are normalized so the mean is 50 and the standard deviation is 10. Then these normalized scores are multiplied by the weights presented in Table A.2. Note that the weights presented in this table are decided by ÖSYM.

[26]It is important to note that to retake the exam students need to wait one more year. So there is a significant opportunity cost associated with retaking.

For Social Science track students, the placement score (Y-ÖSS) is a composite of the track score (ÖSS-SÖZ) and the standardized GPA[27] (AOBP). Students whose placement scores is above 105 points can submit preferences (submit an application) to 2-year college programs, while 120 points are needed to apply to 4-year college programs.

Table A.1 presents the summary statistics by gender for Social Science track students. Examining Social Science track students, the distributions of weighted track scores (ÖSS-SÖZ) as well as normalized GPAs for first time takers by gender are depicted in Figure 1[28]. First, note that females seem to dominate males in terms of GPAs. However, this does not carry over to exams. Males seem to do a bit better than females in the exam at the top of the track score distribution, but considerably worse at the low end. One explanation for this pattern could be greater risk aversion on the part of females which makes them skip questions with a positive expected value. Another could be that males put more effort in the exam as might be expected given the low female labor force participation rates in Turkey.

Figure 1: Score Distributions



3. MULTIPLE-CHOICE EXAM SCORES

We begin by taking a first look at students' scores in the Turkish, Social Science, Math and Science tests. Recall that each section of the exam has 45 questions. The scoring structure results in each multiple of 0.25 between -11.25 and 45 (with the exception of certain numbers above 42) being possible outcomes in an exam.[29] For example, attempting all questions and getting all wrong, results in a score of $-\frac{45}{4} = -11.25$.

[27]The standardized GPA is the GPA normalized by the performance of the school in the university entrance exams which adjusts in effect for different grading practices across schools.

[28]According to the two-sample Kolmogorov-Smirnov test, the distributions of OSS-SOZ score and normalized GPA for males and females are significantly different with the p-value of 0.

[29]Recall that for each question, there are five possible answers; answering correctly gains the student a single point, skipping the question (not giving an answer) gives zero points, but attempting the question and answering incorrectly results in a loss of a quarter point.

Obtaining a particular raw subject score could happen in only one way or in many ways. For example, there is only one way that a student could obtain -11.25 or 45, similarly a score of 42.5 could only have arisen through attempting all questions, getting 43 questions correct and 2 incorrect. On the other hand, a score of 40 has two possible origins: 40 correct and 5 skips, or 41 correct and 4 incorrect. It is impossible to achieve a score of 42.25: the student must have at least 43 questions correct, and at least 3 questions incorrect, which is not possible given there are only 45 questions.

There are 220 possible raw scores one can reach, however if a student attempts all the questions, not skipping any, there are only 46 raw scores that can occur. These are spaced 1.25 points apart, starting at -11.25, and ending at 45 points. The distributions of raw subject scores in Social Science and Turkish for the first time takers by gender, as seen in Figures 2 and 3, have very prominent spikes.[30][31] It is no coincidence that the spikes appear evenly placed; they correspond to the 46 scores that occur after attempting all questions and come from the fact that there is a mass of students, of differing abilities, who answer all the questions. This is an important part of our identification strategy as explained below as we do not have question-by-question data for students.

Math and Science test score distributions for Social Science track students do not exhibit this pattern as most students obtain a score of zero (see Figure 4). Nor do any of the subject score distributions for the Science track students exhibit this pattern of spikes across the entire support of the distribution. These spikes are only there for the top part of the distribution for the science track students consistent with only the very best students attempting all the questions.[32] As Social Science track students do not spend much time on the Science and Math sections of the exam, we assume away the time constraint and restrict our attention to only the Social Science and Turkish sections of the exam for Social Science track students.[33]

4. MODEL

Given the complex relationship between scores, admission outcomes and expected utilities of those outcomes, we do not seek to obtain an explicit utility function (as a function of exam score) in this paper. We assume that the candidate answers each test and each question in isolation.[34] Having utility increase with the score makes sense as a higher score increases the number of programs the student is eligible for, and so gives more options to a student. We do not allow for outcomes in one section of the test to have any bearing on other sections. Expressed alternatively, we do not allow a students perceived performance in previous questions to impact behavior in subsequent questions.[35] Nor do

[30] According to Two-sample Kolmogorov-Smirnov test, the distributions of raw Social Science and Turkish test scores for males and females are significantly different from each other with p-values 0.

[31] Grid lines spaced 1.25 marks apart correspond with these spikes.

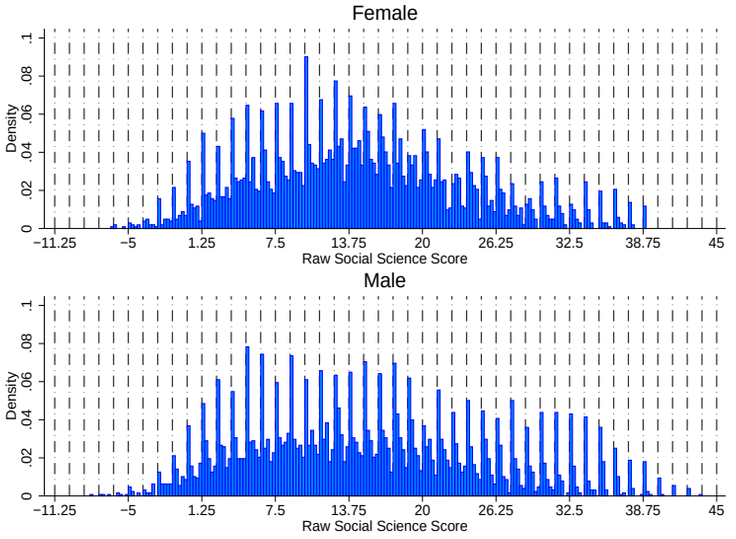
[32] See Figures A.5 and A.6 in Appendix A.1.

[33] In the exam booklet there is a note before the Social Science/Turkish section of the exam that says: "If you want a higher score in ÖSS-SÖZ, it may be better for you to spend more than 90 minutes on verbal section of the exam."

[34] We assume that questions are equally difficult. However, if one has item response data, this assumption can be relaxed (see Online Appendix A.3).

[35] Even if students are able to perceive that they are answering better than expected ex-ante, they are not able to discern if this is due to luck in being asked questions which happen to be well suited to their individual strengths, or if the exam is simply easier than average (implying that scores will be normalized

Figure 2: Distribution of Social Science Test Scores



we allow for any time pressure that results in skipping: students do not skip questions in order to improve their performance in other questions.[36]

Students decide whether to answer a question or skip it, to maximize their expected utility, depending on their probability of answering the question correctly, P_C. We assume that a student makes the decision to skip or answer question by question. We also assume that all questions are of the same difficulty. We make these assumptions as we do not have item response data. Note that as a result, we ignore the possibility that a student who had good signals in previous questions may change his guessing behavior. This may be reasonable in our setting.[37]

A student with an expected score of S will maximize his expected utility and we can write his problem as follows:

$$\max \; I(answer)\left[P_C U(S-1+1)+(1-P_C)U(S-1-d)\right]+I(skip)U(S-1+0)$$

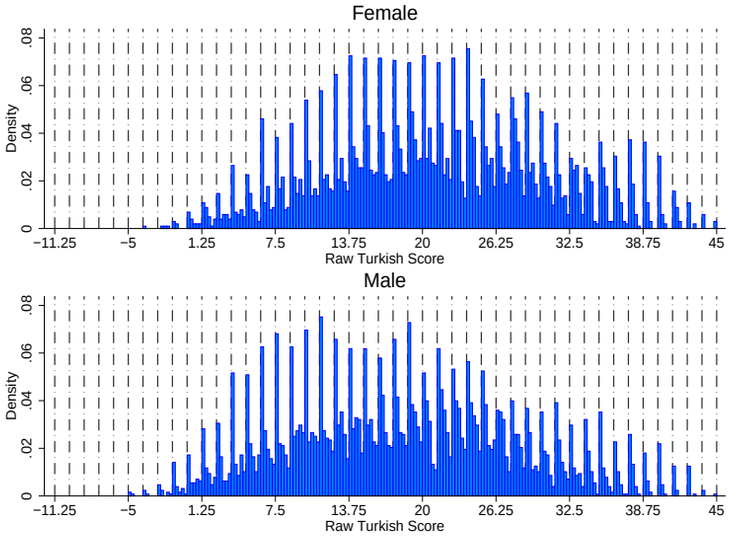
where d is the penalty applied for the wrong answer, and $d \geq 0$.

downwards).

[36]In examining students in the Social Science track we believe this is appropriate, as these students overwhelmingly skip the Science and Math sections of the test, as recommended by examiners, allowing them ample time to focus on the Social Science and Turkish questions.

[37]There is some work that suggests this assumption might be reasonable. The "hot hands" hypothesis argues that in certain environments like sports, a player who has performed well in the recent past will continue to do so. The evidence on this has been mixed in the past. However, a recent paper, see Miller and Sanjurjo (2018), argues that it does exist. We do not allow for the "hot hands" possibility in this paper. The "hot hands" argument is made in settings when success or failure in the recent past is observed. However, in our setup, there is no feedback on how well students are doing while they take the exam. On the other hand, skipping behavior may change as the exam proceeds independent of "hot hands" as students will always know how many questions they have already skipped. Having already skipped many questions the student may decide to take more risk in the hope of reaching a target score.

Figure 3: Distribution of Turkish Test Scores



So, the student will answer the question, if

$$P_C U(S) + (1 - P_C)U(S - 1 - d) > U(S - 1)$$

$$P_C > \frac{U(S-1) - U(S-1-d)}{U(S) - U(S-1-d)} = c$$

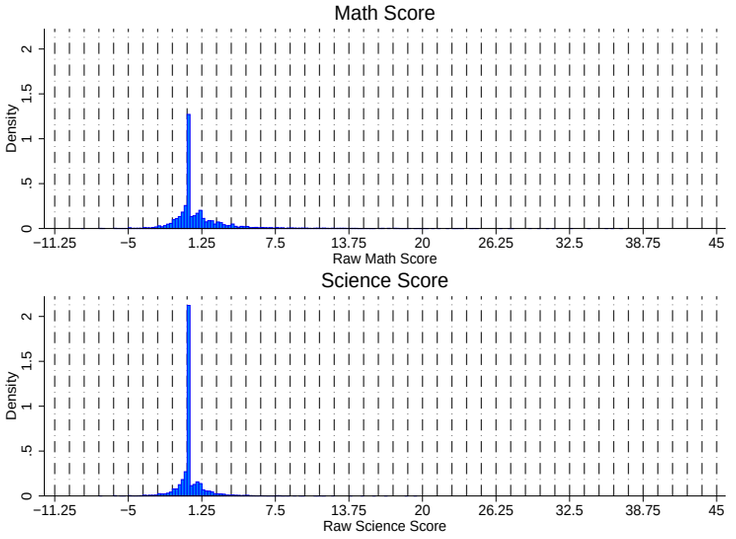
where c will be called the attempt cutoff. If the student's probability of answering the question correctly is above this cutoff, he will answer the question, otherwise he will choose to skip it. c rises with the degree of risk aversion as shown in Appendix A.[38] In our estimation we will allow c to vary by gender and expected score bin. It is worth pointing out that for the most part, we do not make any explicit assumptions about the utility function, $U(.)$, above. We assume a CARA utility function in two places. First in Proposition 2, where we show that the cutoff c is increasing with risk aversion with a CARA specification and second, to pin down a risk aversion level when we change the number of choices in a question in a counterfactual.

In this section, we will construct a model that allows us to structurally estimate these attempt cutoffs as well as the ability distributions of the students (since ability affects the probability of answering the question correctly). We model test taking behavior as follows. When a student approaches a question, he gets a signal for each of the five possible answers. The vector of signals for the question is then transformed into a belief. This belief is the likelihood that an answer is in fact the correct answer. The student then decides whether or not to answer the question, and if so, which answer to choose.

Signals for each of the five answers depend on whether or not the answer is actually correct. Signals for incorrect answers are drawn from a distribution G, where G is Pareto with support $[A_I, \infty)$ and shape parameter $\beta > 0$. Thus, the density of the signal x for

[38] It would also rise if a student lacked confidence, so our estimates of c capture both risk aversion and the confidence level as explained earlier.

Figure 4: Distribution of Math and Science Test Scores of Social Science Track Students



an incorrect answer is $\frac{\beta A_I^\beta}{x^{\beta+1}}$ for $x > A_I$. The mean signal is $\frac{\beta A_I}{\beta-1}$ which is decreasing in β. Signals for correct answers are drawn from a distribution F, where F is Pareto with support $[A_C, \infty)$ and shape parameter equal to $\alpha > 0$, so that the density of the signal is $\frac{\alpha A_C^\alpha}{x^{\alpha+1}}$ for $x > A_C$. The mean signal is $\frac{\alpha A_C}{\alpha-1}$ which is decreasing in α.

ASSUMPTION 1 $A_I = A_C = A$.

This assumption rules out complete certainty that an answer is correct or incorrect.[39]

Suppose that the student observes five signals, given by the following vector:

$$X = (x_1, x_2, x_3, x_4, x_5) \tag{1}$$

where x_i is the signal that the student receives when examining answer i. Using Bayes' rule, the probability that answer i is correct conditional on X, can be expressed as:

(2)

$$\text{Prob}(\text{Answer } i \text{ is correct}|X) = \frac{\text{Prob}(X|\text{Answer i is correct}) \times \text{Prob}(\text{Answer } i \text{ is correct})}{\text{Prob}(X)}$$

Expressing the numerator in terms of the densities of the two distributions, F and G, for the case where $i = 1$:

$$\text{Prob}(X|\text{Answer 1 is correct}) = \frac{\alpha A^\alpha}{x_1^{\alpha+1}} \frac{\beta A^\beta}{x_2^{\beta+1}} \frac{\beta A^\beta}{x_3^{\beta+1}} \frac{\beta A^\beta}{x_4^{\beta+1}} \frac{\beta A^\beta}{x_5^{\beta+1}} \tag{3}$$

[39]For example, assuming that $A_C > A_I$, would mean that it is possible for student to be sure that an answer is wrong.

In essence, this is the density of $F(\cdot)$ at x_1 (as this is conditional on 1 being correct) multiplied by the product of the density of $G(\cdot)$ at the other signals.

It follows, by substituting equation 3 into equation 2, that the probability that answer i is correct, conditional on X, can be expressed as:

$$\text{Prob}(i \text{ is correct}|X) = \frac{\frac{\alpha A^{\alpha}}{x_i^{\alpha+1}} \prod_{j \neq i} \frac{\beta A^{\beta}}{x_j^{\beta+1}}}{\sum_{m=1}^{5} \left(\frac{\alpha A^{\alpha}}{x_m^{\alpha+1}} \prod_{n \neq m} \frac{\beta A^{\beta}}{x_n^{\beta+1}} \right)} \tag{4}$$

where $i, j, m, n \in \{1, ..., 5\}$.

This can be further simplified to:

$$\text{Prob}(i \text{ is correct}|X) = \frac{\frac{1}{x_i^{\alpha+1}} \prod_{j \neq i} \frac{1}{x_j^{\beta+1}}}{\sum_{m=1}^{5} \left(\frac{1}{x_m^{\alpha+1}} \prod_{n \neq m} \frac{1}{x_n^{\beta+1}} \right)} \tag{5}$$

Thus, the choice of A is irrelevant. For this reason we will set it at 1 from here on. Letting $\gamma = \beta - \alpha$, so that $\frac{1}{x_i^{\alpha+1}} = \frac{1}{x_i^{\beta+1}} x_i^{\gamma}$, the expression further simplifies to:

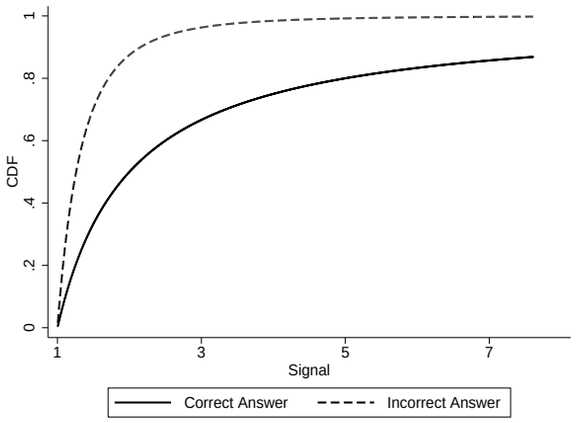
$$\text{Prob}(i \text{ is correct}|X) = \frac{x_i^{\gamma}}{\sum_{m=1}^{5} x_m^{\gamma}} \tag{6}$$

Note that the sum of beliefs for each of the five answers adds up to unity. We assume without loss of generality that $\beta \geq \alpha$, so that the mean signal for the incorrect answer is lower than that for the correct answer. Thus, the higher the signal, x_i, the greater the likelihood that answer i is correct.[40] A higher shape parameter for a Pareto distribution shifts probability mass to the left so that the signals from incorrect answers would generally be smaller. Hence, if we fixed α, a higher γ (i.e., a higher β) would correspond to greater ability. In fact, it is worth emphasizing that it is the difference in the distributions of the signals of correct and incorrect answers that captures ability. Someone who thinks all answers are great is as bad as someone who thinks none of the answers are great: it is the extent to which one can distinguish between the right and the wrong answers that indicates ability. This is why the mean signals mean nothing: it is only the difference in their means that matters. In addition, we assume that the lower bound for signals for both correct and incorrect distributions is the same. Given these assumptions, we can rescale so that the correct answer is drawn from a distribution where $A = 1$ and the shape parameter, α, is also 1, while the signal drawn for an incorrect answer is drawn from a distribution where $A = 1$ and the shape parameter is $\frac{\beta}{\alpha} > 1$. As a result, the structure of a student's signals can be represented by the shape parameter of the incorrect answer: β. A higher value of β draws the the mass of the distribution towards the minimum, $A = 1$, allowing the student to more clearly separate the incorrect signals from the signal given by the correct answer. In other words, higher β students are what would be referred to as high ability

[40] If a student were to draw from distributions with $\beta < \alpha$, smaller signals would be associated with the correct answer and we would reverse our interpretation of the signal.

students. Signal distributions for a student with ability $\beta = 3$ (approximately median) are shown in Figure 5.

Figure 5: Distributions of signals for a student with $\beta = 3$



The effect of a higher β on test outcomes can be decomposed into three effects. First, the correct answer has a higher probability of generating the highest signal. Increasing β shifts the cumulative distribution function (CDF) of the incorrect answers' signals to the left, and the student's best guess (the answer with the highest signal) will be correct more often. Second, when the correct answer actually gives the highest signal, the probability with which the student believes that it comes from the correct answer increases as the weighted sum of the incorrect signals decreases. If the first answer is the correct answer, lowering $\sum_{i=2}^{5} x_i^{\gamma}$ increases the student's belief that answer one is correct.

Finally, there is a subtle effect of β on tests. Students with high ability, i.e. a high value of β, will be more confident in their choices. Even with the same signals, as we increase β, the student's belief that the highest signal comes from the correct answer increases.[41] This is formally stated below:

PROPOSITION 1 *Suppose there are two students: one with ability parameter $\beta = b_1$ and the other with ability parameter $\beta = b_2 > b_1$. Suppose that the two students receive identical signals X for a question. Let $x_{\max} = \max\{x_1, ..., x_5\}$. The student with the higher value of β has a higher belief that $x_{\max}$ is drawn from the correct answer.*

PROOF: The belief is given by $\frac{x_{\max}^{\gamma}}{\sum_{m=1}^{5} x_m^{\gamma}}$. Taking logs, and differentiating with respect to γ, yields the following expression:

$$\frac{d\log(\text{Belief})}{d\gamma} = \log x_{\max} - \frac{x_1^{\gamma}\log x_1 + x_2^{\gamma}\log x_2 + x_3^{\gamma}\log x_3 + x_4^{\gamma}\log x_4 + x_5^{\gamma}\log x_5}{x_1^{\gamma} + x_2^{\gamma} + x_3^{\gamma} + x_4^{\gamma} + x_5^{\gamma}} \tag{7}$$

[41] By signals we refer to the observed vector of x values. To see why ability matters, consider a vector of signals (3,1.2,1.1,1.3,1.2). A high ability student would interpret this as being favorable towards the first answer. A student with no ability, i.e. $\beta = 1$, obtains no information from the signals and can only conclude that all answers have an equal likelihood of being correct.

Since $\log x_{\max} \geq \log x_i$, and $x_i > 0$,

$$\frac{d\text{Belief}}{d\gamma} \geq 0 \tag{8}$$

with the inequality strict unless $x_1 = x_2 = x_3 = x_4 = x_5$. *Q.E.D.*

Once students have observed the signals for each of the five possible answers to the question, they are faced with six possible alternatives: choosing one of the five answers, or skipping the question. Skipping the question does not change their test score, answering correctly increases the score by 1, while answering incorrectly decreases the score by 0.25 points. Note that the expected value of a random guess is $(0.2)(1) - (0.8)(0.25) = 0$.

If a student were to choose an answer, they would choose the one which was most likely to be correct. A slightly higher score is clearly preferred. In this model, the answer which is most likely to be correct is the one with the highest value of x_i. Also, this answer trivially has a probability of being correct (conditional on observed signals and the student's ability) greater than or equal to twenty percent.

As explained, students have a cutoff for the belief below which they will skip the question. If the student believes that the best answer (highest signal) has a probability of being correct greater than this cutoff, he will attempt the question, choosing the best answer. This cutoff lies in the interval $[0.2, 1]$.[42] A higher value for this cutoff implies a higher degree of risk aversion, while a cutoff of 0.2 would be supported by risk neutral preferences. We show next that if we have a CARA utility function then this cutoff is monotonically increasing in the extent of risk aversion.

PROPOSITION 2 *There is a monotonically increasing relationship between the risk aversion parameter, τ in a CARA utility function, and the attempt cutoff, c.*

PROOF: Proof is presented in Appendix A. *Q.E.D.*

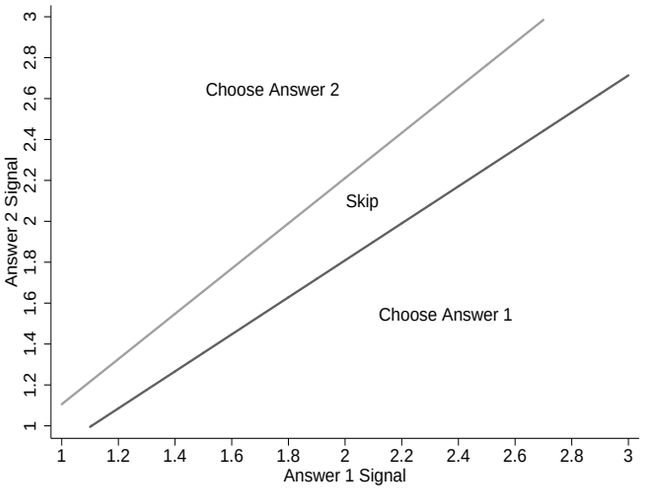
Consider a student with ability parameter β (recall that $\alpha = 1$) and attempt cutoff $c \in (0.2, 1)$. In order to answer a question, with answer i, the signal drawn for answer i, x_i, must satisfy two conditions. First, it must be the highest signal. Second, it must be high enough (given the other signals, and ability β) that the belief that it is correct is greater than c, the cutoff required to attempt the question. A diagram showing choices conditional on signal observations for a simplified two answer setup (with $\beta = 3$ and $c = .55$) is shown in Figure 6. If the signal for j is sufficiently high, then j is selected. In between the two lines, where signals are very close to each other, the best option is to skip the question. This skip region is larger the greater the risk aversion of the agent (the greater the value of c).

5. ESTIMATION STRATEGY

In our model, students' scores depend on students' ability (β) and attempt cutoff, c, which captures attitudes towards risk. In our data set, we observe only the student's score

[42] There will always exist an answer with probability of being correct greater than or equal to 0.2, therefore we do not consider a cutoff below 0.2, as they would result in the same behavior: always attempting the question, never skipping.

Figure 6: Actions for a Question with Two Possible Answers



in each part of the exam, and not the question-by-question outcome, i.e., item response. As a result, we are forced to assume that all questions are of the same difficulty, i.e., the average difficulty. This may bias our estimates of risk aversion as discussed below.

In this section, we explain how we can use our model to estimate the distribution of ability and attempt cutoffs, c, which captures the extent of risk aversion. Estimation of the parameters of interest, the distribution of student ability in Turkish and social science fields $\beta = (\beta_T, \beta_{SS})$ and attempt cutoff c, is conducted separately for each gender. In addition, we recognize that the relationship between ÖSS-SÖZ score and utility is not necessarily constant throughout the range of scores: the degree of risk aversion may be different. In particular, we might expect that students anticipating low scores could be considerably less risk averse, since scores below a cutoff result in the same outcome: an inability to submit preferences/apply to universities. This would result in a jump in the payoff function as students cross the cutoff score. For this reason we allow attempt cutoffs to vary by gender, and allow them to depend on the interval in which the student's predicted Social Science track score (ÖSS-SÖZ) lies, for example 120-130. This predicted score in effect proxies for ability. We cannot use students' actual exam scores as a proxy for ability as these are endogenous objects that are affected by students' risk taking behavior in the exam. Therefore, we predict students' scores by using their observable characteristics. Specifically, GPA (adjusted for school quality)[43], education level of both parents, income levels/monthly income of parents, preparation on the four subject areas, and the school type. We run an OLS regression separately for males and females first time takers in the Social Science track, and use the results to predict track (ÖSS-SÖZ) scores for each student (see Table A.3 in Online Appendix A.1).

[43]To adjust for school quality, we adjust the GPA of student within a school based on the performance of the school in the exam. We observe normalized GPA for each student, which is converted to a ranking within the school. As we observe the mean and variance of exam scores for each school, we can easily convert the GPA to a measure that reflects the quality of the school.

5.1. *Estimation*

We divide students into groups, according to gender, and the range into which their predicted track score (ÖSS-SÖZ) lies: $(0, 90)$, $[90, 100)$, $[100, 110)$, $[110, 120)$, $[120, 130)$, $[130, 140)$, and $[140, \infty)$.[44] These groups do not contain equal numbers of students, but do contain at least 100 students.[45] For each group, we examine the two subjects (Social Science and Turkish) jointly as we allow correlation in the ability of a student in the two. We assume that students in each score group have a common attempt cutoff, c, and draw from the joint distribution of ability $(\beta_{Turkish}, \beta_{SocialScience})$. The ability of each student in subject $k \in (T, SS)$ is given by $1 + e^{\psi_k}$, where (ψ_T, ψ_{SS}) are distributed normally with mean $\mu = (\mu_T, \mu_{SS})$ and covariance matrix Σ.[46] This ensures that each student has an ability in both subjects greater than 1, and results in a log normal distribution (shifted 1 unit to the right).[47] It also allows for abilities in the two subjects to be correlated, as would typically be the case.[48]

Under the assumptions made, the probability of obtaining each score is approximated through simulation. For student n, we take a draw from $N(\mu, \Sigma)$ and label the vector as ψ_n. From ψ_n, we find $(\beta_T, \beta_{SS}) = (1 + e^{\psi_s(1)}, 1 + e^{\psi_s(2)})$, the student's ability vector. As we now have (β_T, β_{SS}, c) for student n, we can generate the simulated student's test outcome, namely the Turkish score and Social Science score.

In order to estimate the relevant parameters for the group (cutoff, means of ψ_T, ψ_{SS}, variances of ψ_T, ψ_{SS} and correlation between ψ_T and ψ_{SS}), we use simulated method of moments. For every group we postulate a cutoff, the mean of ψ_T, ψ_{SS}, the variance of ψ_T, ψ_{SS} and correlation between ψ_T and ψ_{SS}. We make 100 draws for each student in the group and construct the relevant moments for the group. These moments are the mean scores in the two subjects, the variance of these scores, the correlation between the scores in the two subjects, and the intensity of the spikes in the two subjects. The difference between the mass of students with scores corresponding to attempting all questions (i.e. $45, 43.75$,..,-11.25) and the mass of students with scores corresponding to skipping a single question (i.e. $44, 42.75$,...,-11) is what we mean by the intensity of the spikes.[49] If the spikes are very prominent, this difference will be large; if they are non-existent, this difference will be minimal. In a given exam, for each such pair, we find this difference and take its sum to get the overall measure of the spike intensity. This gives us two more moments to match.[50]

[44] Most of the students in this bin has predicted scores between 140 and 150.

[45] With the exception of females in the lowest expected score range.

[46] In practice, correlation coefficients ρ were obtained rather than covariance, to assist the minimization procedure and for interpretation. The covariance obtained is therefore $cov(T, SS) = \rho\sigma_T\sigma_{SS}$.

[47] It can be shown that the likelihood of answering correctly increases approximately linearly with respect to the log of ability, so that a log-normally distributed ability would generate the roughly normal score distribution observed.

[48] Within the Social Science track as a whole, the scores in the Turkish and Social Science sections are highly correlated, the correlation coefficient is 0.78 with the p-value 0.

[49] There are many ways to get a particular score. For example, 35 can be reached by correctly answering 35 and skipping 10, correctly answering 36 and skipping 9 (4 incorrect) or correctly answering 37 with 8 incorrect. This multiplicity is not generating the spikes. As seen in Figure 6 below, the prevalence of spikes is clearly driven by risk aversion.

[50] There are alternative ways to measure the intensity of the spikes. It is also possible to define the spike intensity for each section - Turkish and Social Science- separately. We do not need to do so as we have a single cutoff that defines risk aversion so that a single measure suffices for identification.

We compare simulated test score moments to those observed in the data and choose the parameters that minimize the objective function. Accordingly, the estimates of the vector $\theta^g = (c^g, \mu^g, \Sigma^g)$, cutoff c and ability distribution parameters for each group g, denoted by the vector $\hat{\theta}^g = \left(\hat{c}^g, \hat{\mu}^g, \hat{\Sigma}^g\right)$, are estimated by minimizing the distance between the simulated moments, $\hat{m}(g)$, and the observed moments, $m(g)$.

$$\hat{\theta}^g = \arg\min_{\theta}(\hat{m}(g) - m(g))' W_T^{-1} (\hat{m}(g) - m(g)) \tag{9}$$

where W_T is the weighting matrix. As usual, we begin by using the identity matrix as the weighting matrix thereby obtaining an estimate of the parameters of each group that is consistent and asymptotically normal. Applying the two step procedure, (Hansen (1982), Gourieroux and Monfort (1997), Duffie and Singleton (1993)) this estimate is used to generate a weighting matrix. Using the new weighting matrix, the procedure is repeated which improves the efficiency of the estimate.

For a given c, the means of scores help pin down the means of the ability distributions that students are drawing from, and the variances/covariances of scores help pin down the variances of the ability distributions and the correlation between ability draws. Identification of the attempt cutoff, c, is achieved through matching the intensity of the spikes in the score distributions for Turkish and Social Science. If students are less risk averse then they will tend to not skip, ceteris paribus. Thus, at low values of c, almost all of the probability mass of a given student's distribution will be located on scores corresponding to attempting all questions and resulting in spikes. As c increases, students become more and more likely to skip some questions, resulting in more mass lying on scores unreachable by attempting all questions so that spikes can no longer be seen. Similarly, as the question difficulty rises, fewer students will attempt all questions and again the intensity of the spikes will lessen.

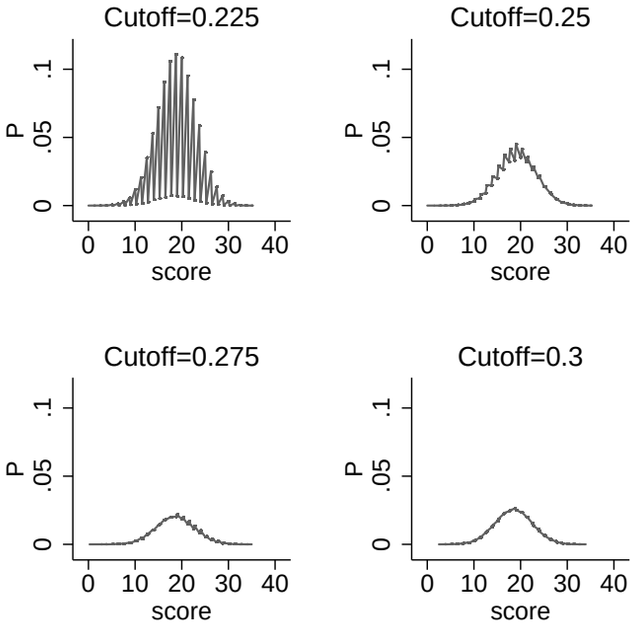
This is illustrated in Figure 7[51], where the simulated score distribution for a student (with a fixed, approximately median, ability of $\beta = 3$) is shown for various cutoff levels. A cutoff of $c = 0.225$ puts virtually all of the mass of the score distribution on values that correspond to students attempting all questions. As the attempt cutoff increases to 0.3, the spikes all but disappear as very few attempt all questions. The relationship between the intensity of the spikes and the attempt cutoff is not constant. As we increase ability, given a cutoff c, the intensity of the spikes increases. This makes sense as high ability agents are more likely to distinguish the right answer from the wrong one and answer all questions for any given cutoff. While low ability students are not likely to have an answer with a belief above the attempt cutoff, this becomes increasingly common as ability rises. This is shown in Figure 8[52], where the attempt cutoff is set to 0.25 [53]. While there are multiple ways through which a particular score may be obtained, this number does not drive the spikes we see in the score distribution. Figure A.4 in Appendix A.1 depicts the number of

[51] Only the set of scores corresponding to attempting all questions and the set of scores corresponding to skipping a single question are shown as identification is based on the relative probabilities of the two sets of scores.

[52] Only the set of scores corresponding to attempting all questions and the set of scores corresponding to skipping a single question are shown as identification is based on the relative probabilities of the two sets of scores.

[53] Ability ranges from approximately the 20^{th} to the 80^{th} percentiles, as estimated

Figure 7: Distribution of scores resulting from various cutoff levels



combinations (the density) that give rise to each score (which is on the horizontal axis). Comparing this figure to the density of Social Science and Turkish test scores (Figures 2 and 3) shows they look nothing alike. This suggests that skipping behavior is driving the latter, not the difference in the number of ways a score can be attained.

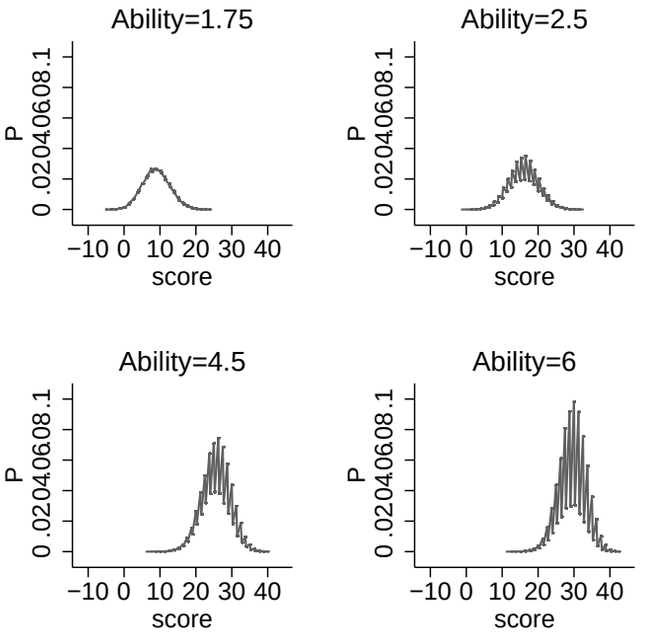
The parameters of the distribution of the ability of a group of students, (μ_T, μ_{SS}) and Σ, are identified by the distribution of scores. An increase in the mean parameter μ_T moves the Turkish score distribution to the right, increasing the mean, while an increase in the variance parameter σ_T^2 increases the variance of the Turkish score distribution. This is due to a strong relationship between ability and exam scores. Similarly with the Social Science section. Finally, the correlation between Turkish and Social Science ability draws is reflected in the correlation of scores.[54]

6. RESULTS

In Figure 9, we display the estimated cutoffs (the belief below which a student skips) graphically. Table 1 presents the estimation results in column (1) and (2) for females and males, respectively, and column (3) presents the t-statistics for the difference between

[54]In the absence of item response data we cannot distinguish between ability and question difficulty as, given risk aversion, both result in more skipping. This is why we assume all questions are of equal difficulty. In the general model which can be applied with item response data we can recover both ability and risk aversion of each individual, as well as question difficulties, item by item. The intuition is as follows. The fraction who get a question correct will fall with difficulty. The best students will answer even the most difficult questions correctly. Thus, as question difficulty rises, the curve linking ability and the probability of getting the question correct swings down from its anchor at the highest ability (who get all questions correct). This will separate ability and question difficulty. Risk aversion will be pinned down by skipping behavior.

Figure 8: Distribution of scores resulting from different ability levels



attempt cutoffs of males and females. Attempt cutoffs for males and females are significantly different in all bins except the lowest and highest. Two facts are apparent. Males tend to have lower attempt cutoffs, especially for students whose predicted score is above the threshold that allows them to submit a preference list.[55] This is in line with the literature as discussed previously. Secondly, the cutoff is systematically lower in the predicted score ranges below 120. This matches what we know about the payoff structure. For low scores, students should be much less risk averse since any score below 105 will not allow the student to submit preferences for any school, and any score below 120 will not permit the student to submit preferences for four year college programs. Above 120, the cutoff remains relatively high[56] and seems to rise with the predicted score bin consistent with increasing risk aversion.[57]

[55]The estimates for second time takers presented in Appendix A.1 Table A.5 show the same patterns: the cutoff rises with the expected score and cutoffs are slightly higher for women suggesting greater risk aversion for women and those with a higher expected score bin. Table A.6 presents the difference between the estimated cutoffs between second and first time takers. We see estimated c being higher for second time takers in low expected score bins (consistent with higher risk aversion and/or under confidence) but lower for second time takers in high expected score bins (consistent with lower risk aversion and/or over confidence for these students).

[56]Cutoffs for the top students are approximately 0.26, which has meaning that these students will only answer a question if they are at least 26% sure of their answer. Significantly more than the 20% likelihood of a random guess.

[57]We also re-estimate our model, allowing the cutoff to differ between two test subjects. These results are presented in Appendix Table A.4. The cutoff for Social Science is slightly higher than for Turkish (the difference is in the second or third decimal place). One reason for this might be that students tend to find Turkish easier and to be overconfident in Turkish as this part of the test evaluates reading comprehension, sentence structure and grammar. As Turkish is the mother tongue, students may naturally find Turkish

TABLE 1
ESTIMATES OF RISK AVERSION CUTOFF

	(1)	(2)	(3)
Expected Score Interval	Female	Male	t-stat for (female cutoff-male cutoff)=0
(0,90)	0.214	0.215	-0.506
	(0.003)	(0.002)	
[90,100)	0.23	0.227	2.425
	(0.001)	(0.001)	
[100,110)	0.239	0.235	2.864
	(0.001)	(0.001)	
[110,120)	0.253	0.249	2.056
	(0.002)	(0.001)	
[120,130)	0.266	0.258	2.378
	(0.003)	(0.002)	
[130,140)	0.274	0.262	2.503
	(0.004)	(0.003)	
[140,inf)	0.271	0.264	1.141
	(0.005)	(0.003)	

Notes: Standard errors are reported in parentheses.

Recall that we cannot separate risk aversion from under-confidence. However, since other work does, we can use their estimates to get some idea of the extent to which our estimates would differ from those that captured pure risk aversion. In a setting on exam taking Iriberri and Rey-Biel (2021) find that one standard deviation increase in the overconfidence leads to 0.060143 (0.0137*4.39) standard deviation less omission, however, one standard deviation increases in risk aversion increase omission by 0.7575 (0.202*3.75) standard deviation. Using their numbers we find that about 8% of the difference in skipping behavior (0.0137*4.39 standard deviations versus 0.202*3.75 standard deviations) comes from differences in confidence. So our estimates of differences in risk aversion should be shrunk by 8% to account for the differences in confidence.

Figures 10 and 11 show the simulated score distributions compared to observed distributions for the various groups. The figure clearly consists of spikes for attempting all questions. To help with visualization, these are presented in different grid lines corresponding to scores which could be obtained by skipping no questions. While the estimation procedure was designed only to match subgroups of the sample, the entire simulated distribution fits the data relatively well overall. It is worth noting that estimation methods which grouped students based on actual track (ÖSS-SÖZ) score did better here.

Estimates of the parameters governing the distribution of ability for each group are presented in Table 2. Recall that ability is parametrized as $(1 + e^{\psi}, 1 + e^{\psi})$, where $\psi \sim N(\mu, \Sigma)$. The means and variances of the components of ψ in each group are presented.

As we estimate the distributions for students in the Social Science track, differences in ability distributions could come from selection into this track as well as differences given selection. For example, if the track was seen as friendly to females in some way, it might

easier and have a false sense of confidence in this exam. Alternatively, differences in the degree of question difficulty heterogeneity may result in unequal bias.

Figure 9: Estimates of Attempt Cutoffs: Social Science Track

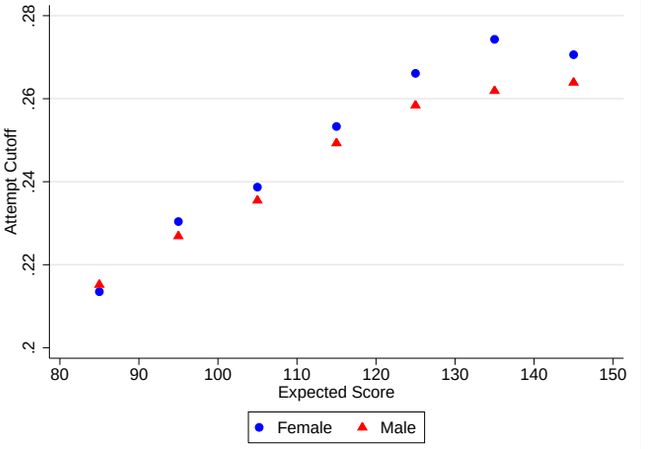


Figure 10: Data vs Simulated Score Distribution: Social Science

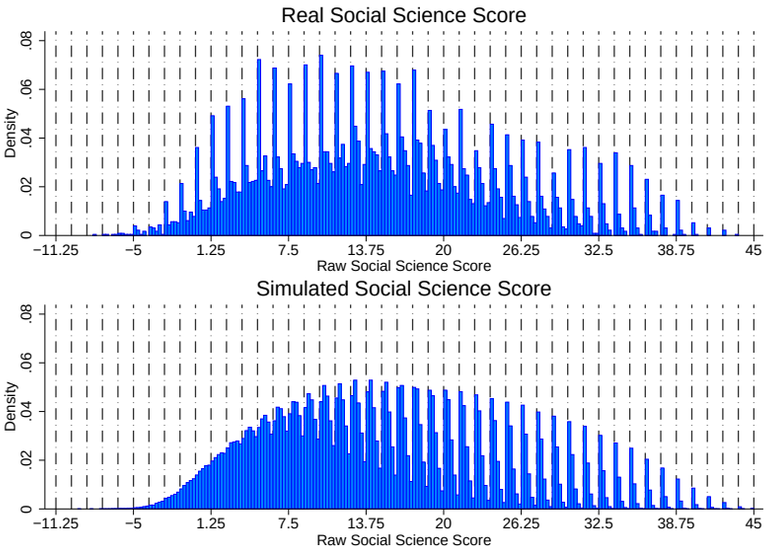
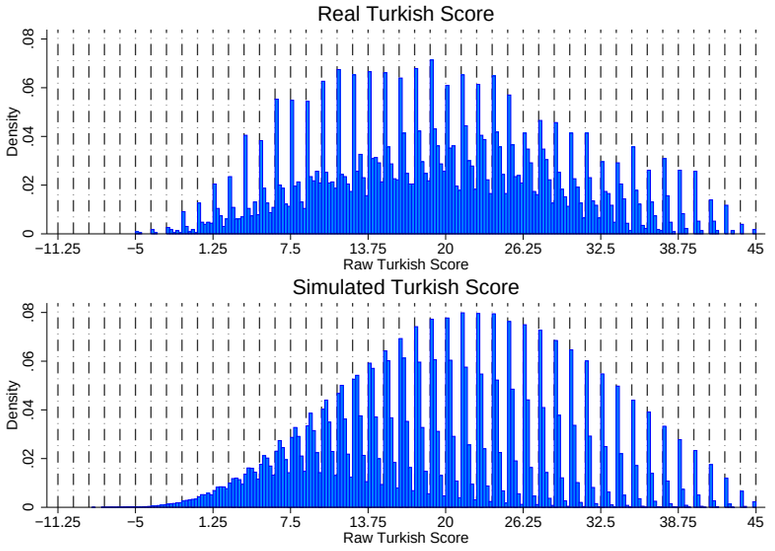


Figure 11: Data vs Simulated Score Distribution: Turkish



attract more females, especially weak ones, with better females going into less female-friendly tracks. With this qualification, we see that females tend to have higher ability in Turkish, but slightly lower ability in Social Science, when compared to males in the corresponding group.[58] This is consistent with males having a comparative and absolute advantage in Social Science and is consistent with findings in the literature that females have the advantage in Language skills (See Lynn (1992)).

In addition, we observe that males tend to have higher variance in their distribution of ability as shown in Table 2. In fact, the variance is greater in all deciles.[59] The correlation between ability in Turkish and Social Science seems to be higher in each decile for females, as seen in Table 3. This would tend to give females an advantage in terms of being at the top: in order to gain admission students must perform well in both Turkish and Social Science. It would also explain the higher variance for males.

We assume that all questions are at the same level of difficulty in both the model and estimation procedure. This assumption is necessary as we do not observe item level responses. Would this create biases in our estimation? We check for this using a simulation based approach. We first simulated a model where questions had differing difficulties and estimated (on this generated data) a model where questions had the same level of difficulty. The details of this can be found in Appendix A.4. We find that the cutoffs are biased upwards by this assumption. The reasoning behind this is straightforward. In the data, there tend to be relatively few skips for each student. If a student skips on average k questions when there is no heterogeneity, when heterogeneity is introduced (keeping average difficulty the same), the student tends to skip more than k questions as hard questions are much more likely to be skipped than average questions. In other words, the probability of skipping is convex in difficulty. As a result, the average number of skips at the mean

[58]The estimated ability distributions for the two sections of the test are depicted in Figure A.2 in Appendix A.1.

[59]The estimated ability distributions for Turkish and Social Sciences by gender reflect this higher variance as in Figures A.3.

TABLE 2

ESTIMATES OF ABILITY DISTRIBUTION PARAMETERS

Social Science Test				
	Female		**Male**	
	μ	σ	μ	σ
(0,90)	-1.101	0.778	-1.377	0.975
	(0.208)	(0.18)	(0.178)	(0.118)
[90,100)	-0.647	0.713	-0.624	0.799
	(0.045)	(0.038)	(0.037)	(0.033)
[100,110)	-0.077	0.591	0.07	0.705
	(0.024)	(0.02)	(0.024)	(0.021)
[110,120)	0.491	0.562	0.693	0.652
	(0.023)	(0.022)	(0.025)	(0.021)
[120,130)	1.022	0.490	1.28	0.572
	(0.033)	(0.029)	(0.033)	(0.028)
[130,140)	1.491	0.434	1.72	0.573
	(0.048)	(0.047)	(0.051)	(0.044)
[140,inf)	1.914	0.427	2.215	0.344
	(0.061)	(0.083)	(0.054)	(0.079)
Turkish Test				
	Female		**Male**	
	μ	σ	μ	σ
(0,90)	-0.379	0.72	-0.853	0.773
	(0.163)	(0.133)	(0.112)	(0.092)
[90,100)	0.079	0.581	-0.146	0.668
	(0.032)	(0.027)	(0.028)	(0.025)
[100,110)	0.577	0.534	0.398	0.61
	(0.020)	(0.018)	(0.021)	(0.018)
[110,120)	1.100	0.525	0.926	0.569
	(0.022)	(0.020)	(0.022)	(0.019)
[120,130)	1.669	0.469	1.431	0.526
	(0.031)	(0.031)	(0.031)	(0.028)
[130,140)	2.226	0.418	1.976	0.557
	(0.048)	(0.048)	(0.050)	(0.048)
[140,∞)	2.906	0.607	2.618	0.556
	(0.091)	(0.098)	(0.077)	(0.089)

Notes: Standard errors are reported in parentheses.

TABLE 3
ESTIMATES OF CORRELATION BETWEEN LOGS OF TURKISH AND SOCIAL SCIENCE ABILITY

	Female	**Male**
(0,90)	1.000	0.783
	n/a	(0.108)
[90,100)	0.920	0.848
	(0.035)	(0.029)
[100,110)	0.948	0.888
	(0.024)	(0.021)
[110,120)	0.921	0.928
	(0.027)	(0.02)
[120,130)	0.990	0.900
	(0.042)	(0.036)
[130,140)	1.000	0.948
	n/a	(0.054)
[140,∞)	0.862	0.841
	(0.144)	(0.211)

Notes: Standard errors are reported in parentheses. Because of the small number of observations in the female sample, (0,90) and [130,140) predicted score groups, standard errors cannot be calculated.

level of difficulty is less than the average number of skips when there is heterogeneity in question difficulty. Another way of saying this is as follows. Intuitively, suppose we see K skips in the data which comes for a setting where questions differ in difficulty. If we assume all questions have the same difficulty, these skips give us one estimate of c. However, if questions are heterogeneous in difficulty (with the same average difficulty) the same number of skips would pin down a lower c (as a lower c reduces skipping behavior and heterogeneous questions raise the number of skips). In other words, by assuming questions have the same difficulty, we would over estimate c. Though our estimates are likely overestimates, this will not impact our main conclusions, namely that the difference in risk aversion by gender has small effects on scores, and that removing negative marking will reduce the accuracy of the exam in identifying ability as students will guess more often.

We also run the model for the Language test in the Language track and the Turkish exam in the Turkish-Math track as a robustness check. These are presented in Appendix A.2. It is reassuring to note that the estimates look a lot like those above, despite the raw data looking quite different.

7. COUNTERFACTUALS

Having recovered the parameters regarding risk aversion exhibited by students in the multiple-choice tests, in addition to estimates regarding the distribution of ability (as measured by β for each subject, the parameter in the Pareto distribution that governs dispersion of signals), we are now able to perform counterfactual experiments.

In these experiments, we compare outcomes of a number of testing regimes, and student behaviors. For example, how would exam outcomes differ if all students attempted (answered) every question, as would happen if the penalty for answering incorrectly were

removed. This is relevant because it is fully feasible to change the testing regime, and the regime may well affect outcomes. Our focus is on two points. First, we look at the gender gap, defined as the over-representation of males at the top of the score distribution. This comes both from ability differences and from differences in behavior. In particular, we will quantify the impact of risk aversion differences on test outcomes as the test format changes. Second, we look at the effectiveness of a test as the format varies. Effectiveness is defined as the extent to which performance matches abilities. The rationale behind penalties is to reduce the amount of random guessing, reducing noise in scores and improving effectiveness. The downside is that as females seem to be significantly more risk averse than males, this accentuates the gender gap. Our focus is to understand the extent of this trade-off.

For this reason we consider seven possible regimes in our counterfactual experiments. These are:

1. The baseline model, as estimated in the previous section.
2. All students attempt all questions. Removing penalties for wrong answers which would make all students attempt all questions.[60] If students are risk neutral, they should also attempt all questions as long as the expected value of guessing is zero.
3. Risk preferences of females are altered, so that the cutoff used by a woman in predicted ÖSS-SÖZ score interval k is changed to that used by a man in predicted ÖSS-SÖZ score interval k (labeled as "Equal Cutoffs" in the figures).
   Note that the second regime eliminates the effects of risk aversion as well as differences across gender. The third regime keeps risk aversion, but eliminates the gender differences in risk aversion. While this is not feasible to perform in practice, we can use the counterfactual exercise to quantify the effect of gender differences in risk aversion in the current system.
4. Each question has only four answers to choose from, with the penalty for an incorrect answer adjusted to keep the expected value of guessing with no information at zero. This will increase the impact of risk aversion and accentuate the gender gap and hinder the effectiveness of the exam. Why? Reducing the number of choices makes the gamble involved in answering have higher stakes. This should exacerbate the effect of different risk preferences across the genders. In the default regime, there are five answers, with a single point for correct answers and a quarter point lost for incorrect answers. This results in an expected gain of zero from a random guess; accordingly, we set the penalty equal to one third of a point in the four answer scenario, resulting in a random guess having an expected gain of zero. As a result, the cutoffs for attempting a question must be different. To convert cutoffs from the five answer case, we assume a CARA utility function, and solve for the risk aversion parameter that generates a given cutoff. This is repeated for each group. We then convert the risk aversion parameter to a cutoff in the four answer case.[61] Note that having four answers instead of five, and increasing the penalty accordingly, increases the variances of scores for a given student even in the absence of risk

[60] In this case, scores would need to be rescaled to reflect the absence of such a penalty: instead of ranging from -11.25 to 45, they would range from 0 to 45.

[61] For example, a cutoff of 0.240 in the five answer case implies risk aversion coefficient of 0.383 (CARA utility), which results in a cutoff of 0.300 in the four answer case.

aversion.[62]

5. The penalty for answering incorrectly is increased from 0.25 points to 1 point. This will accentuate the gender gap but increase effectiveness as guessing is reduced. This counterfactual is designed to elicit more skipping from students and to amplify the impact of differences in risk preference across the genders. As in the four-answer counterfactual, cutoffs are translated into implied CARA parameters and new cutoffs are obtained for both counterfactuals.
6. The number of questions in each section is doubled, from 45 to 90. This will improve the effectiveness of the exam and could increase the gender gap if males are more prevalent at higher abilities. This allows us to place results in the context of a more precise exam: increasing the number of questions increases the ability of the exam to distinguish students based on ability.

For each of the six possible regimes, we find the resulting distributions of scores for the entire sample of (first time takers) students in the Social Science track.

We simulate the model using the parameters estimated,[63] generating scores in the Turkish and Social Science section, adding the two to generate an exam score for each student.[64] We then segment students into bins by using the total score. The bins are constructed such that five percent of students are in each bin, so that we observe the 5% of students who perform the worst, the 5% who perform the best etc.[65]

7.1. *Distinguishing Between Student Ability*

We first examine the effect of the different regimes on the ability of the exam to select the most capable students. To do so we look into the relationship between exam score percentile in the data and (average log) ability in the two subjects, Turkish and Social Science. For each score bin, we have estimated the mean of the ability distribution students draw from. For each counterfactual, we generate a score for each individual and a distribution of scores. As each individual has an ability draw in the baseline, we can obtain the mean ability of the students in each bin of the score distribution for the counterfactual. If this mean ability is below that in the baseline for the lowest bins and above it for the highest bins, i.e., the difference is negative for the lowest bins, rises monotonically, ending up being positive for the highest bins, the counterfactual is better able to sort on the basis of ability. Figures 12 and 13 show the difference in mean ability between the counterfactual of interest and the baseline model.

The Turkish and Social Science graphs show the same patterns. We see that the "Attempt All", "Equal Cutoffs", and "Four Answers" regimes all show very little difference in terms of the quality of admitted students in Figures 12 and 13. The "Attempt All" regime basically makes the cutoff 0.2 for all bins and genders. The "Equal Cutoffs" regime makes

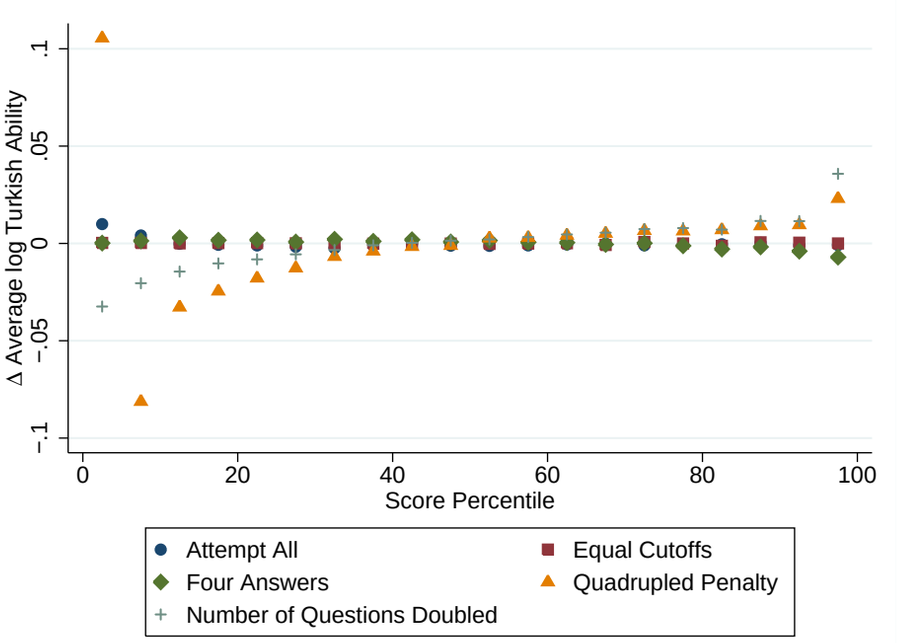
[62]The standard deviation of the points earned for a single question is, for a student of (approximately median) ability $\beta = 3$, 0.66 (four answers) vs 0.62 (five answers) i.e. scores are more precise when there are five answers than when there are four answers. For a student of ability $\beta = 6$ (approximately the top 10%) the standard deviation is 0.58 vs 0.56.

[63]1000 students were simulated for each student observed in the data.

[64]We did not simulate scores from math and science as the majority of students skipped these sections, and scores of those who attempted were close to zero.

[65]As seen in Figures A.1a and A.1b in the Appendix, the exams do sort by ability as higher score percentiles are associated with higher average ability in all the regimes studied.

Figure 12: Turkish Ability (Δ vs baseline)



female cutoffs equal to male cutoffs in each score bin. The "Four Answers" regime basically raises the cutoff for both genders. However, as the cutoffs do not vary much by bins or gender, this is translated into small impacts in terms of the ability to discriminate.

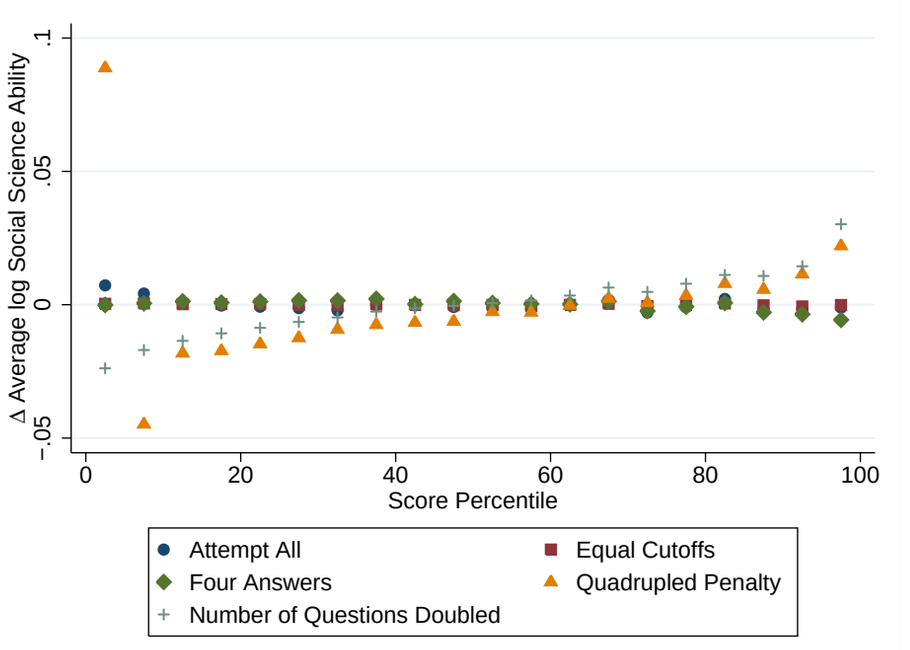
Higher penalties clearly do a better job at sorting, especially at the bottom. Average abilities under these regimes are lower than the baseline on the left (more accurately identifying weak students)[66] and higher than the baseline on the right (more accurately identifying strong students). The reason for the higher effectiveness of the high penalty regime is simple. It strongly discourages guessing, and more so when uninformed. This reduces the variance in the score distributions of an individual student, resulting in a cleaner signal. The downside is that differences in risk aversion by gender will have more of an impact.

Both greater penalties for wrong answers and more questions improve the ability of the exam to sort students. How do they compare to each other? The impact of the increased penalties on average abilities of combined score quantiles is most evident for the top quantiles. Note that the top 13.5% roughly get admission in the Social Science track. We find that an additional 25 questions (70 in total) must be asked in each section in order for the baseline model to have a comparable admitted class, to the 45 question, quadrupled penalty version.[67]

[66] With the exception of the quadruple penalty regime for the lowest ventile. Examining more carefully, the lowest 5% actually has a higher average ability than the second lowest 5% (see Figures A.1a and A.1b). This is not due to any risk aversion differences (the pattern remains even if all students are given the same cutoff of 0.25). The explanation is simple: The bottom 5% tends to consist of students who attempted questions and had bad luck. Since attempting is related to a higher ability we observe this interesting pattern.

[67] Alternatively, if the penalty were quadrupled, the number of questions in each section could be reduced to only 27 yet would retain equivalent validity.

Figure 13: Social Science Ability (Δ vs baseline)



7.2. *Impacts on Gender Bias*

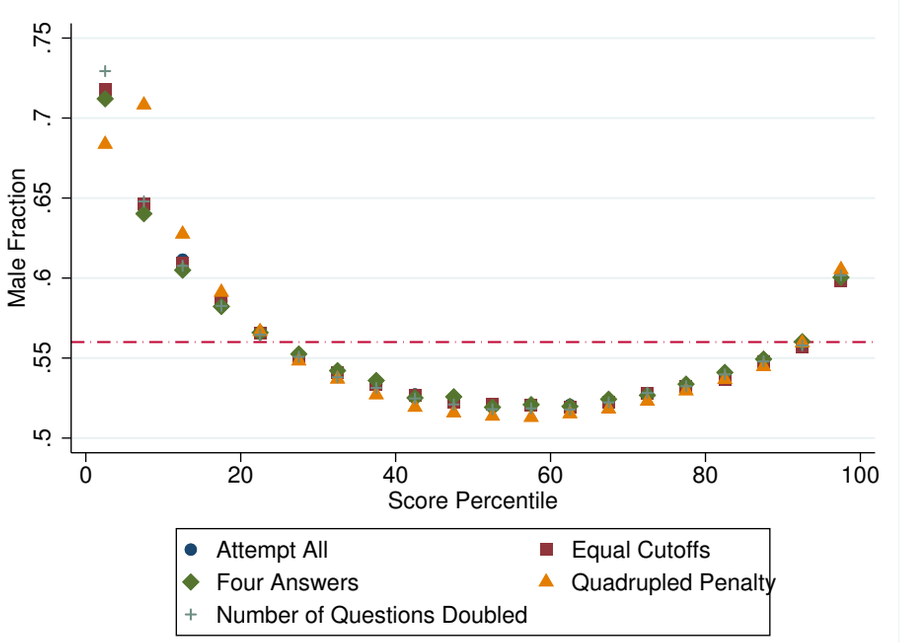
Finally, we examine the impact of the various regimes on the fraction of males in the different score percentiles. In particular, we want to see if there is any effect on the (over)-representation of males in the top scoring percentiles: the gender gap. Lower risk aversion of males raises the presence of males at the top and the bottom. Consequently, we would expect the male fraction in the data to be U shaped. We do not draw this in Figure 14, which shows the male fraction in each of the counterfactual scenarios. We omit the baseline male fraction as it is hard to distinguish from those in the counterfactuals. The average fraction male in the Social Studies track is drawn as a horizontal line at 0.56.

In Figure 15, we depict the difference in the male fraction in each counterfactual and the baseline. These differences are all close to zero. There is a minimal reduction in the gender gap if we were to eliminate skipping ("Attempt All"), or eliminate risk aversion differences ("Equal Cutoffs"). The impact on the gender gap of the remaining counterfactuals, while slightly higher, remains small. Only the quadrupled penalty increases the fraction male at the bottom bins consistent with males having a lower risk aversion and guessing more. This is more damaging with a quadrupled penalty when ability is low.

Why do risk aversion differences seem to matter so little? There are two reasons. Firstly, there is a relatively low chance that a student has a belief lying between 0.23 and 0.25, for a given question. Secondly, if the belief does lie in this region, the expected gain from answering (and hence that from having a low cutoff) is at most 0.0625 points. Even when the penalty is raised, leading to more skipping behavior, the total effect on allocations is minor. Essentially, differences in choices made due to skipping behavior are not common, and when they do arise have small consequences. Intuitively, this is like saying that while ordering at a restaurant, the best option is usually clear, and when it is not, the choice made has little consequence.

To make this point, we present some simple simulations. Table 4 gives the fraction

Figure 14: Male Fraction Counterfactuals



correct, incorrect and skipped, as well as the expected scores, and the gains in expected score from being risk neutral at three ability levels with the ability parameter being drawn from a distribution with $\beta = 2,\ 3$ or 4. Risk aversion changes from $.2$ to $.325$.

We see that an increase in risk aversion raises the gains from being risk neutral at each ability level. However, these gains increase as ability falls. The small impact of risk aversion, especially for students with high ability, is clear when examining Table 4. A student with a cutoff of 0.275 with ability $\beta = 4$ (approximately top 25%) has an expected score 0.09 lower than a risk neutral student of the same ability. The impact on students in the top 10% is even smaller.

Figure 15: Male Fraction (Δ vs baseline)

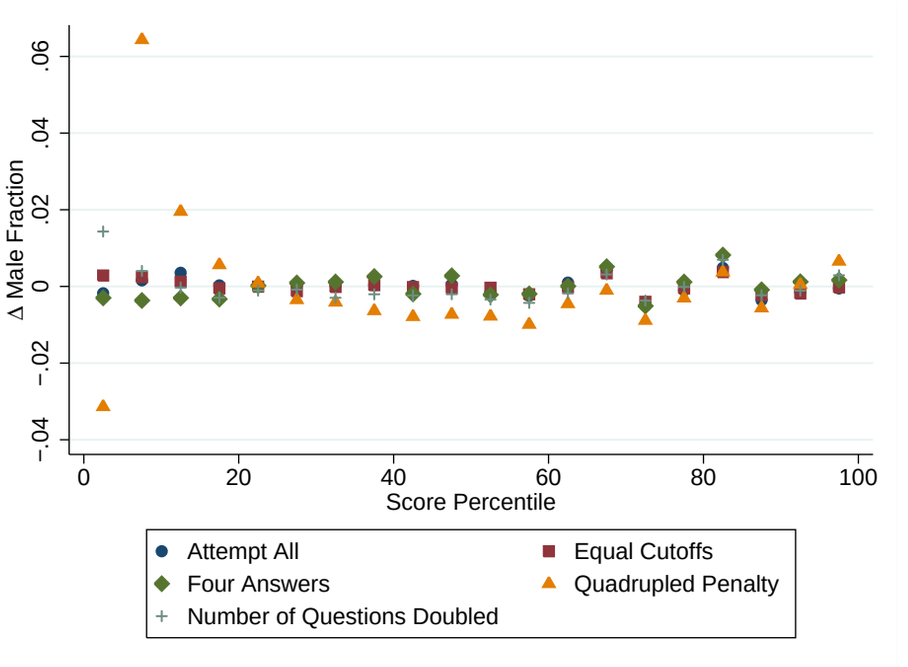


TABLE 4

QUESTION OUTCOMES FOR VARIOUS PARAMETER VALUES: PROBABILITIES OF SKIPPING (S), BEING CORRECT (C), BEING INCORRECT (I), EXPECTED SCORE OUT OF 45, AND THE REDUCTION IN EXPECTED SCORE AS COMPARED TO A RISK NEUTRAL STUDENT OF THE SAME ABILITY

β	Cutoff	Prob(S)	Prob(C)	Prob(I)	Expected Score	Loss vs Risk Neutral
2	0.2	0	0.405	0.595	11.57	-
2	0.225	0.012	0.403	0.585	11.57	0.00
2	0.25	0.085	0.386	0.529	11.43	0.14
2	0.275	0.192	0.359	0.449	11.12	0.45
2	0.3	0.303	0.328	0.370	10.58	0.99
2	0.325	0.403	0.297	0.300	9.99	1.58
3	0.2	0	0.535	0.465	18.86	-
3	0.225	0.003	0.534	0.463	18.86	0.00
3	0.25	0.030	0.528	0.442	18.81	0.05
3	0.275	0.081	0.515	0.404	18.63	0.23
3	0.3	0.143	0.498	0.360	18.36	0.50
3	0.325	0.208	0.478	0.315	17.96	0.90
4	0.2	0	0.619	0.381	23.58	-
4	0.225	0.001	0.619	0.380	23.58	0.00
4	0.25	0.017	0.616	0.368	23.58	0.00
4	0.275	0.049	0.608	0.344	23.49	0.09
4	0.3	0.091	0.596	0.314	23.27	0.31
4	0.325	0.137	0.581	0.281	23.00	0.58

8. CONCLUSIONS

In this paper, we construct a structural model of a student's decision to attempt/skip a question in a multiple-choice exam in the presence of a guessing penalty. Different from the Item Response Theory (IRT) models such as the Rasch Model, our model can deal with skips. We identify the risk aversion parameter of the model by using the idea that the lack of risk aversion makes certain scores more likely which creates spikes in the score distribution. We use our model and estimates to better understand two questions.

First, women tend to do worse in high stakes exams, a pattern that is prevalent in many settings. One reason postulated has been greater risk aversion or under-confidence on the part of females which makes them less wiling to guess and reduces their performance in multiple choice exams. We find that while females do act in a more risk-averse manner, the impact of this is relatively limited in terms of performance and the prevalence of females in the group admitted to the university. Thus, we need to look elsewhere to quantitatively explain why women do worse than men in such settings. A hypothesis worth exploring is the extent to which this arises from females having different preferences and their preferred schools being less selective and so needing lower scores.[68]

Second, we explore the efficacy of negative marking in more accurately ranking students on the basis of ability. Negative marking raises the effectiveness of the exam in terms of being able to distinguish between students on the basis of ability. It also creates a bias against female students if they are more risk averse. We explore this tradeoff and find that negative marking has a considerable impact on the effectiveness of the exam: a penalty of -1 is similar to doubling the number of questions. Moreover, it does so with a minimal impact on gender bias.

The model developed in this paper can also be extended in a number of ways. These include modeling the effect of time limits in exams and incorporating a cost of effort (that could be individual specific) that rises with time spent on the exam.[69]. In future work, we hope to move in these directions.

APPENDIX A: PROOFS

PROOF OF THE PROPOSITION 2: :

$$\begin{aligned} c(\tau) &= \frac{\exp(-(S-1+d)\tau)-\exp(-(S-1)\tau)}{\exp(-(S-1+d)\tau)-\exp(-S\tau)} = \frac{\exp(-(S-1)\tau)[\exp(d\tau)-1]}{\exp(-(S-1)\tau)[\exp(d\tau)-\exp(-\tau)]} \\ &= \frac{\exp(d\tau)-1}{\exp(d\tau)-\exp(-\tau)} \end{aligned}$$

$$\begin{aligned} \frac{\partial c(\tau)}{\partial \tau} &= \frac{\partial\left[\frac{\exp(d\tau)-1}{\exp(d\tau)-\exp(-\tau)}\right]}{\partial \tau} \qquad (10) \\ &= \frac{d\exp(d\tau)\left[\exp(d\tau)-\exp(-\tau)\right]-\left[\exp(d\tau)-1\right]\left[d\exp(d\tau)+\exp(-\tau)\right]}{\left[\exp(d\tau)-\exp(-\tau)\right]^2} \\ &= \frac{-d\exp(d\tau-\tau)-\exp(d\tau-\tau)+d\exp(d\tau)+\exp(-\tau)}{\left[\exp(d\tau)-\exp(-\tau)\right]^2} \\ &= \frac{-(d+1)\exp(d\tau-\tau)+d\exp(d\tau)+\exp(-\tau)}{\left[\exp(d\tau)-\exp(-\tau)\right]^2} \qquad (11) \end{aligned}$$

[68] We explore this in Akyol, Krishna, and Lychagin (2022).

[69] For these patterns in PISA data, see for example Akyol, Krishna, and Wang (2021).

As the denominator of the expression is positive, it is enough to show that the nominator is positive. We want to show that

$$d\exp(d\tau) + \exp(-\tau) > (d+1)\exp(d\tau - \tau)$$

Divide both sides of the equation by $(d+1)\exp(d\tau - \tau)$

$$\begin{aligned} \frac{d\exp(d\tau) + \exp(-\tau)}{(d+1)\exp(d\tau - \tau)} &> 1 \\ \frac{d}{d+1}\exp(\tau) + \frac{1}{d+1}\exp(-d\tau) &> 1 \end{aligned}$$

Since the exponential function is a strictly convex function, the following inequality holds

$$\frac{d}{d+1}\exp(\tau) + \frac{1}{d+1}\exp(-d\tau) > \exp(\frac{d}{d+1}\tau - \frac{1}{d+1}d\tau) = \exp(0) = 1$$

Q.E.D.

A.1. *Additional Tables and Figures*

Table A.1: Summary Statistics

	Female			Male		
Variable	Obs	Mean	Std. Dev.	Obs	Mean	Std. Dev.
Variable	Obs	Mean	Std. Dev.	Obs	Mean	Std. Dev.
ÖSS-SÖZ score	3,984	110.064	15.414	4,928	108.654	16.673
Normalized High School GPA	3,984	48.663	8.342	4,928	46.122	7.866
Raw Turkish Score	3,984	20.926	9.372	4,928	18.326	9.729
Raw Social Science Score	3,984	14.479	8.890	4,928	15.446	10.037
Raw Math Score	3,984	0.856	3.007	4,928	0.898	2.831
Raw Science Score	3,984	0.084	1.221	4,928	0.152	1.292
Education level of Dad						
Primary or less	3,984	0.522		4,928	0.586	
Middle/High School	3,984	0.308		4,928	0.268	
2-year higher education	3,984	0.026		4,928	0.020	
College/Master/PhD	3,984	0.047		4,928	0.043	
Missing	3,984	0.097		4,928	0.082	
Education level of Mom						
Primary or less	3,984	0.767		4,928	0.818	
Middle/High School	3,984	0.143		4,928	0.115	
2-year higher education	3,984	0.008		4,928	0.007	
College/Master/PhD	3,984	0.012		4,928	0.011	
Missing	3,984	0.070		4,928	0.048	
Prep School Expenditure						
No prep school	3,977	0.337		4,909	0.337	
Scholarship	3,977	0.010		4,909	0.010	
<1000 TL	3,977	0.205		4,909	0.213	
1000-2000 TL	3,977	0.075		4,909	0.058	
>2000 TL	3,977	0.016		4,909	0.014	
Missing	3,977	0.356		4,909	0.368	
Income Level						
<250 TL	3,913	0.427		4,857	0.485	
250-500 TL	3,913	0.409		4,857	0.372	
500-750 TL	3,913	0.104		4,857	0.088	
750-1000 TL	3,913	0.033		4,857	0.030	
1000-1500 TL	3,913	0.015		4,857	0.014	
1500-2000 TL	3,913	0.007		4,857	0.005	
>2000 TL	3,913	0.005		4,857	0.005	
Time Spent in Math Preparation						
<100 hours	3,984	0.117		4,928	0.117	
100-200 hours	3,984	0.078		4,928	0.068	
>200 hours	3,984	0.022		4,928	0.014	
Time Spent in Science Preparation						
<100 hours	3,984	0.079		4,928	0.072	

(continued on next page)

	Female			Male		
Variable	Obs	Mean	Std. Dev.	Obs	Mean	Std. Dev.
100-200 hours	3,984	0.017		4,928	0.013	
>200 hours	3,984	0.004		4,928	0.004	
Time Spent in Turkish Preparation						
<100 hours	3,984	0.075		4,928	0.083	
100-200 hours	3,984	0.104		4,928	0.103	
>200 hours	3,984	0.046		4,928	0.039	
Time Spent in Social Sci. Preparation						
<100 hours	3,984	0.078		4,928	0.085	
100-200 hours	3,984	0.100		4,928	0.093	
>200 hours	3,984	0.064		4,928	0.065	

TABLE A.2

TEST WEIGHTS

	Math	Science	Turkish	Social Science	Language
Science Track (ÖSS-SAY)	1.8	1.8	0.4	0.4	0
Social Science Track (ÖSS-SÖZ)	0.4	0.4	1.8	1.8	0
Turkish-Math Track (ÖSS-EA)	0.8	0.4	0.8	0.3	0
Language Track (ÖSS-DIL)	0	0	0.4	0.4	1.8

Table A.3: Regression to Predict ÖSS-SÖZ Score

Variable	Male	Female
Normalized High School GPA	0.760***	0.643***
	(0.013)	(0.013)
Income Level (base: <250 TL)		
250-500 TL	0.359	0.373
	(0.369)	(0.373)
500-750 TL	-0.116	-0.361
	(0.622)	(0.595)
750-1000 TL	-0.191	-0.589
	(1.003)	(0.985)
1000-1500 TL	0.444	-0.016
	(1.463)	(1.443)
1500-2000 TL	4.522*	0.129
	(2.541)	(2.146)
>2000 TL	4.178	5.668**
	(2.568)	(2.485)
Education level of Mom (base: Primary or less)		
Middle/High School	0.387	-0.310
	(0.566)	(0.522)
2-year higher education	-1.978	3.023
	(2.073)	(1.874)
College/Master/PhD	-0.314	2.369
	(1.786)	(1.829)
Missing	-0.094	-0.591
	(0.997)	(1.058)
Education level of Dad (base: Primary or less)		
Middle/High School	1.292***	1.059***
	(0.405)	(0.399)
2-year higher education	0.166	-0.124
	(1.197)	(1.070)
College/Master/PhD	2.287**	0.942
	(0.927)	(0.946)
Missing	-0.643	0.756
	(0.785)	(0.923)
Time Spent in Math Preparation		
<100 hours	2.504***	3.206***
	(0.948)	(1.081)
100-200 hours	3.876***	4.742***

(continued on next page)

Variable	Male	Female
	(1.111)	(1.197)
>200 hours	5.573***	1.071
	(1.872)	(1.718)
Time Spent in Science Preparation		
<100 hours	1.981**	0.918
	(0.783)	(0.760)
100-200 hours	-0.477	-1.736
	(1.597)	(1.430)
>200 hours	-5.832**	2.519
	(2.933)	(2.897)
Time Spent in Turkish Preparation		
<100 hours	-0.113	-1.416
	(1.202)	(1.437)
100-200 hours	-0.539	-1.740
	(1.183)	(1.345)
>200 hours	-1.090	-0.094
	(1.467)	(1.552)
Time Spent in Social Science Preparation		
<100 hours	-1.079	1.954*
	(1.015)	(1.176)
100-200 hours	-0.430	2.141**
	(1.016)	(1.091)
>200 hours	1.037	1.622
	(1.094)	(1.170)
Prep School Expenditure (base: No prep school)		
Scholarship	4.386**	5.521***
	(1.744)	(1.691)
<1000 TL	5.548***	3.947***
	(0.599)	(0.613)
1000-2000 TL	5.167***	5.080***
	(0.899)	(0.836)
>2000 TL	6.247***	5.371***
	(1.655)	(1.572)
Missing	-0.512	-0.219
	(0.387)	(0.395)
Constant	28.128***	37.628***
	(1.400)	(1.434)
School Type Control	Yes	Yes
Observations	4,823	3,894
R2	0.550	0.566

Figure A.1: Counterfactual Relationship between Ability and Scores

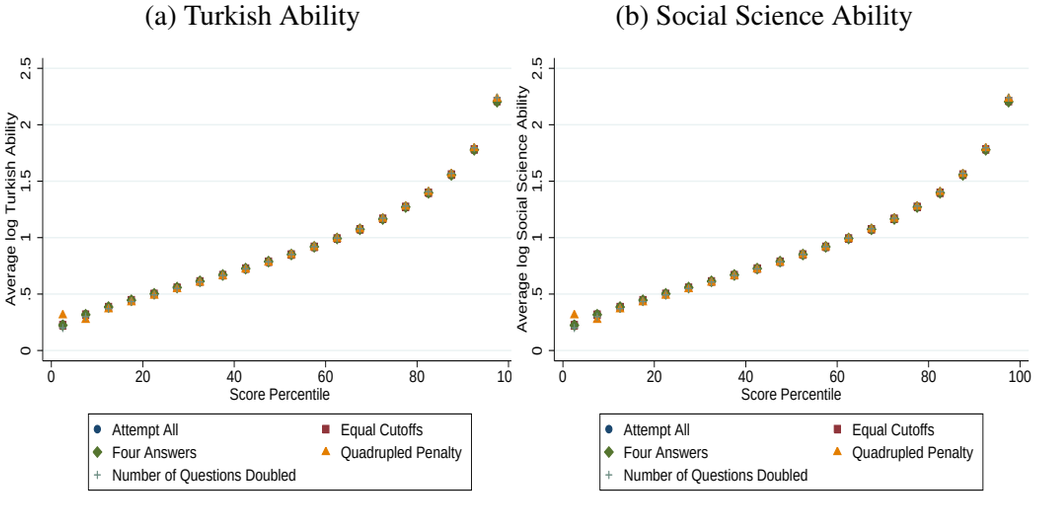


Figure A.2: Distributions of Social Science and Turkish Ability

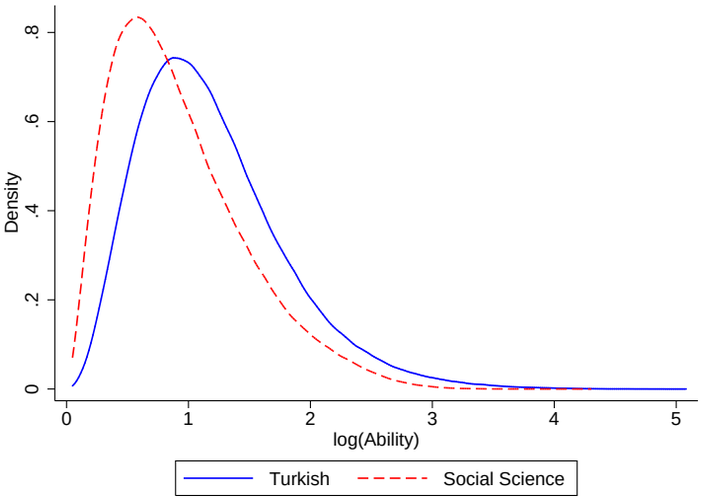


TABLE A.4

ESTIMATES OF RISK AVERSION CUTOFF SEPARATELY FOR SOCIAL SCIENCE AND TURKISH

	Female		Male		t-stat for (female cutoff-male cutoff)=0	
Expected Score Interval	Turkish	Soc. Sci.	Turkish	Soc. Sci.	Turkish	Soc. Sci.
(0,90)	0.21	0.218	0.214	0.216	-0.639	0.674
	(0.003)	(0.004)	(0.005)	(0.002)		
[90,100)	0.219	0.231	0.224	0.228	-3.149	2.001
	(0.001)	(0.001)	(0.001)	(0.001)		
[100,110)	0.23	0.24	0.232	0.237	-0.876	2.231
	(0.001)	(0.001)	(0.001)	(0.001)		
[110,120)	0.244	0.255	0.245	0.252	-0.275	0.987
	(0.002)	(0.002)	(0.001)	(0.002)		
[120,130)	0.253	0.269	0.252	0.268	0.331	0.193
	(0.002)	(0.003)	(0.001)	(0.004)		
[130,140)	0.262	0.277	0.256	0.269	1.136	1.142
	(0.004)	(0.005)	(0.003)	(0.004)		
[140,inf)	0.265	0.285	0.259	0.269	0.827	1.493
	(0.005)	(0.009)	(0.004)	(0.005)		

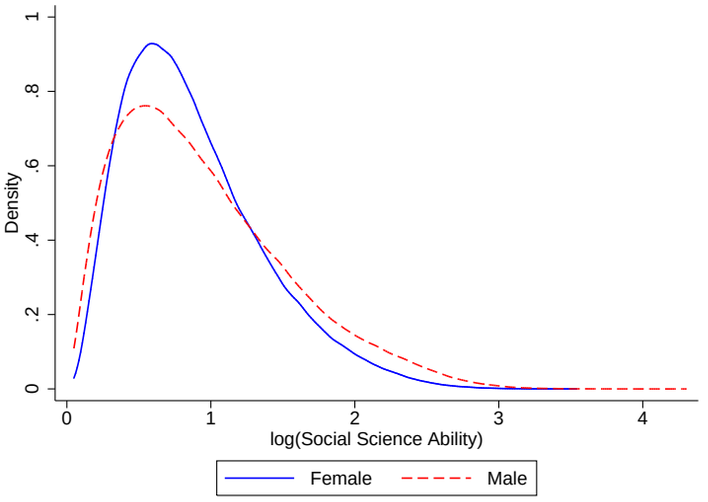
TABLE A.5

ESTIMATES OF RISK AVERSION CUTOFFS FOR SECOND TIME TAKERS

	Female	Male	t-stat for (female cutoff-male cutoff)=0
Expected Score Interval	(1)	(2)	(3)
[90,100)	0.233	0.233	-0.152
	(0.004)	(0.002)	
[100,110)	0.244	0.239	3.036
	(0.001)	(0.001)	
[110,120)	0.253	0.246	5.528
	(0.001)	(0.001)	
[120,130)	0.263	0.252	5.908
	(0.002)	(0.001)	
[130,140)	0.264	0.257	2.248
	(0.003)	(0.002)	
[140,inf)	0.265	0.256	1.526
	(0.006)	(0.003)	

Figure A.3: Distributions of Ability by Gender

(a) Social Science



(b) Turkish

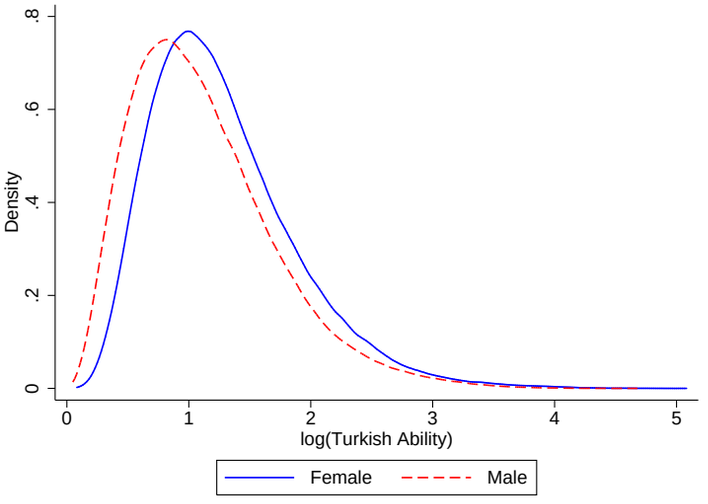


TABLE A.6

DIFFERENCE OF RISK AVERSION CUTOFF BETWEEN SECOND AND FIRST TIME TAKERS

Cutoff 2nd-Cutoff 1st

Females		Males	
Difference	t-stat	Difference	t-stat
0.019	4.117	0.006	2.044
0.013	6.99	0.004	2.627
0.014	9.497	-0.003	-2.587
0.009	3.873	-0.007	-4.049
-0.002	-0.468	-0.005	-1.92
-0.01	-1.299	-0.008	-2.016

Figure A.4: Distribution of Possible Raw Scores

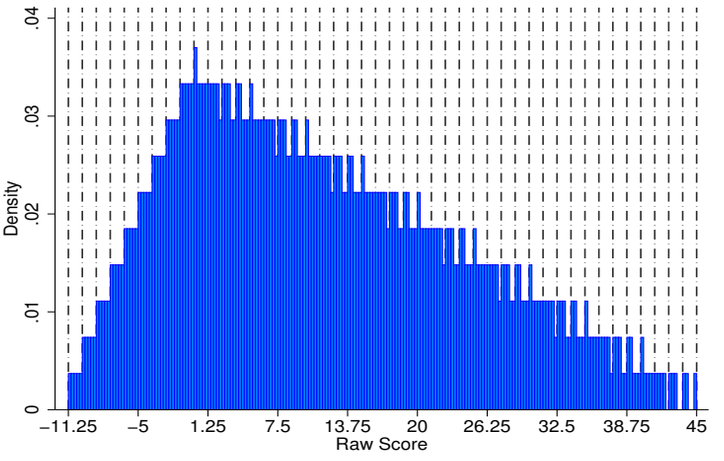


Figure A.5: Distribution of Math Test Scores of Science Track Student

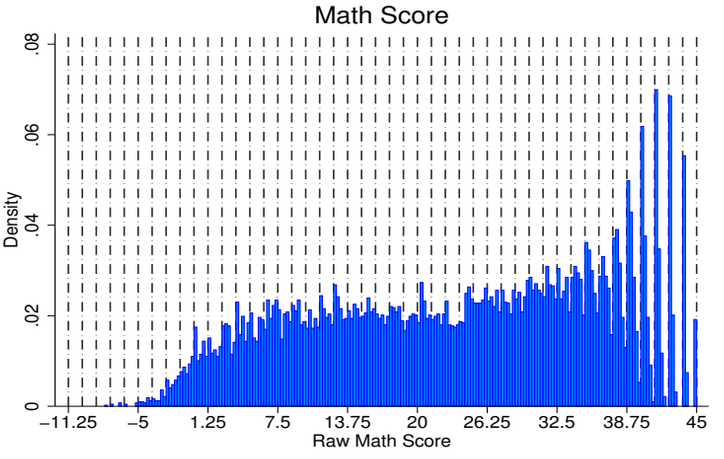
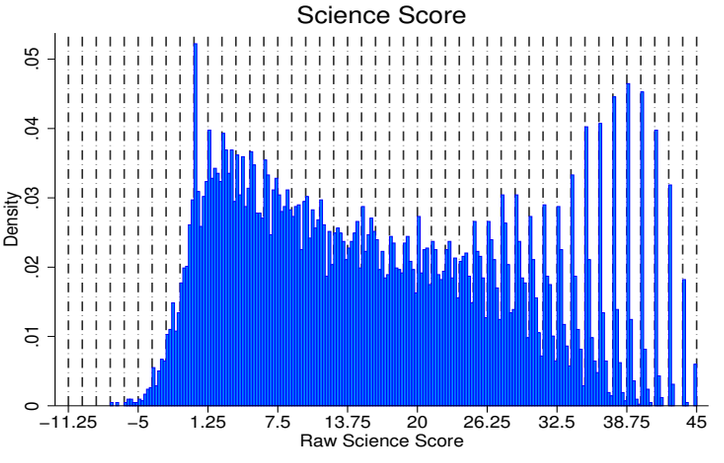


Figure A.6: Distribution of Science Test Scores of Science Track Student



A.2. *Other Tracks*

So far we have focused on the Social Science track (ÖSS-SÖZ). Our approach can be used for subjects where there is partial knowledge such as the Turkish component in the Turkish-Math and Language tracks. As seen in Table 1, the Turkish-Math track (ÖSS-EA) also places high emphasis on the Turkish section of the exam, a section which is well described by the model. The model also applies to the Language section of the Language track (ÖSS-DIL), which as would be expected, accounts for a large part of the student's admission score. We do not use our model on the Math and Science tests. This is because of the limited presence of spikes which are a key part of our identification strategy. The lack of spikes we hypothesize, comes from the questions being of a different type. Science and Math problems have to be worked out and if done correctly, this eliminates all but one answer. As a result, there is a lack of partial knowledge: either the signal is fully informative, or there is no information.

There are some differences between the Language exam and the regular exam. First, the Language exam is held separately, some time after the regular (Science, Math, Turkish and Social Science sections) exam. In addition to this, students are able to view the correct answers following the regular exam. This would give the Language track students information regarding their score in the regular exam. As the Social Science and Turkish sections contribute to the Language track score (albeit a small contribution) this information is relevant. Secondly, although the scoring system is the same for each question (1 point for correct, -0.25 for incorrect, 5 possibilities and the option to skip), there are in total 100 questions in the Language exam. As previously, we only observe the total section score.

We estimate the model for the Turkish-Math track students, examining the Turkish section only. We then estimate the model for the Language track students, examining the Language exam only.

A.2.1. *Estimation*

Estimation follows that in the main body of the paper. After separating first attempt students into predicted ÖSS score bins by gender, we use simulated method of moments to obtain the distribution of ability, and the attempt cutoff for the group. As we are only examining one subject, the ability distribution is univariate. Moments to be matched are analogous to before.

To obtain the predicted score bins, we run a regression between score and observable characteristics, and use predicted values. While the Turkish Math section binning process is the same as before, the Language track is slightly different. As students are able to see the general test questions and correct answers after the exam, it is reasonable to expect students to accurately determine their score from the Turkish and Social Science sections of the exam, at least to a reasonable degree. We therefore use the students' actual performance in the general test when predicting their score in the Language exam.

A.2.2. *Data*

Focusing on students making their first attempt, we obtain 7972 female and 7919 male students for the Turkish Math track, and we obtain 9280 female and 3681 male students for the Language track[70].

The aggregate score patterns can be seen in Figures A.7 and A.8.[71] While the Turkish exam section of the Turkish-Math track students looks relatively similar to previous histograms, the Language track students illustrate a much different pattern. This is due to the large number of questions of the Language section: 100 compared to 45 in other sections. As a result, the medium ability students will tend to skip enough questions for the spikes to diminish greatly. While in the other exam sections there were 45 opportunities for a student to skip a question, in the Language section there are more than double the amount of chances to skip. It follows that there will be more skipped questions, which have the effect of reducing the intensity of the spikes, especially in the middle of the distribution.[72] Estimating the model for data showing a very different aggregate pattern also serves as a robustness check.

[70]There were three different foreign language options, each with their own exam. We chose to focus on the English language students as they were the vast majority. Sample sizes are those for the English language track.

[71]Gridlines are 1.25 points apart in Figure A.7 and 2.5 apart in Figure A.8.

[72]This was also the location where the spikes were least intense in the social science and Turkish sections

Figure A.7: Turkish Score Distribution of Turkish Math Track Students

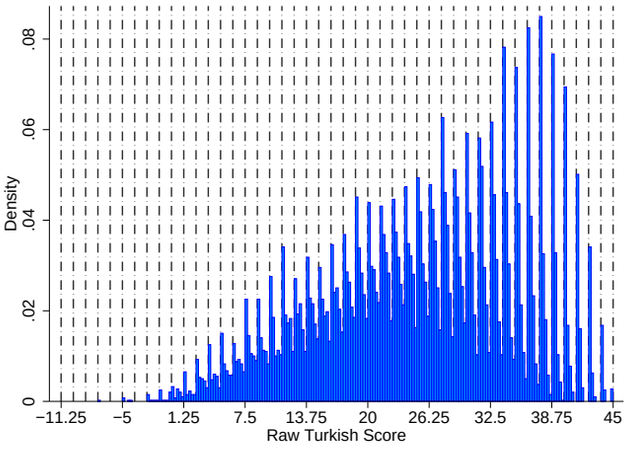
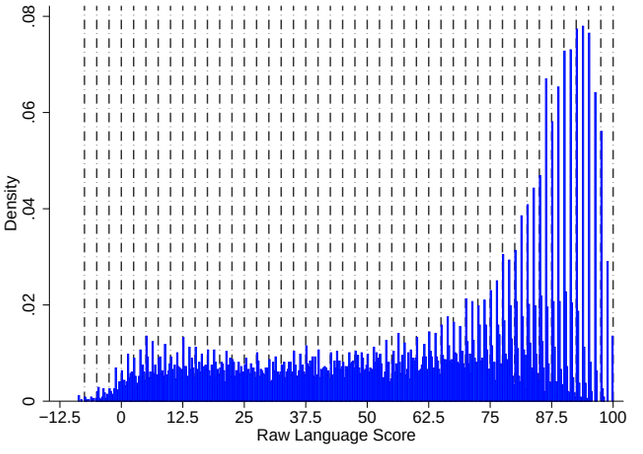


Figure A.8: Language Score Distribution of Language Track Students



A.2.3. *Results*

Estimates of ability distributions for the different groups. For the sake of brevity, only the attempt cutoffs are presented.

Figure A.9: Estimates of Attempt Cutoffs: Turkish Math Track

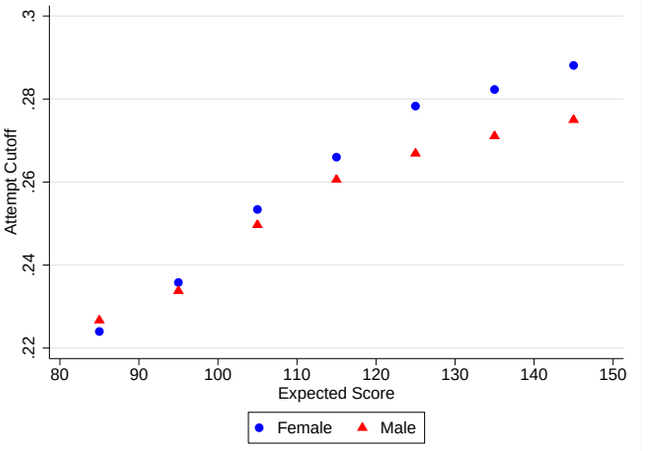
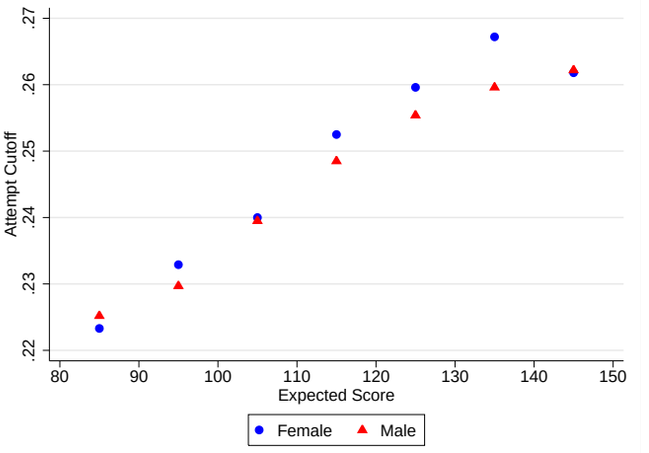


Figure A.10: Estimates of Attempt Cutoffs: Language Track (English)



As before, there are two important patterns. First, as seen in Figures A.9 and A.10, the cutoff increases as we move from students who expect to perform poorly to students who expect to perform well. This is in line with expectations, given the payoff structure: students are less risk averse for low scores as they are below the cutoff for applying. Secondly, males tend to have lower cutoffs than females, i.e., they are less risk averse, and this difference tends to be significant only in higher score bins. Another important observation is that these cutoff patterns are very similar to those observed in the Social Science track. Even the Language track, where the data exhibited very different patterns, has a similar pattern of risk aversion, providing further support for the model and empirical approach.

Note the magnitude of and patterns in the cutoffs, and the degree of differences between male and female students, are relatively similar across tracks. In all three tracks, cutoffs rise with score and males are less risk averse.

A.3. *Extension of Model to Item Response Data*

The data used to estimate students' behavior in our paper is limited, showing only the raw scores in each section. Therefore, we have assumed that students within a group are identically and independently

distributed in their test-taking characteristics and that questions have the same difficulty level. However, it is possible to say more about individual students and individual questions with a dataset that includes item-level scores of students.

In this section, we present an extended version of the model presented in Section 4 that can be used to provide estimates taking advantage of item-by-item responses. We, then, apply this model to a small sample of students taking a mock exam in a "dersane" (schools that prepare students for the exam).

A.3.1. *The Extended Model*

As in Section 4, students approach each question by observing signals for each possible answer. Signals for incorrect and correct answers are draw from Pareto distributions, with identical support but different shape parameters. The student interprets the signals rationally, and finds probabilities that each answer is correct. The student will either answer the question, choosing the answer with the highest likelihood of being correct, or skip the question. Again, the choice is determined by comparing the likelihood of success to some cutoff $c \in [0.2, 1]$.[73]

In contrast to the model presented in Section 4, here we allow questions to vary in difficulty. Some questions will be difficult, and in the context of the model, this will be a result of having similar signal distributions for the correct and incorrect answers. Some questions will be easy which means students will tend to have very different distributions for the correct and incorrect answers. As before, students are heterogeneous in their abilities in the different exam sections.

We need to introduce the following parameters to extend the model: question difficulty $q > 0$ and student ability $s > 0$. Both correct and incorrect answers generate signals drawn from a Pareto distribution with scale parameter 1.[74] The distribution for the correct answer has shape parameter q, and the distribution for incorrect answers has shape parameter $q + s$. As shown previously, it is the ratio of the shape parameters that determines the student's ability to distinguish the correct answer from the incorrect answers. Thus, a student with ability s considering a question with difficulty q will have an effective ability (comparable to β previously) of $k = \frac{q+s}{q} > 1$. As with β, higher values of k are associated with a higher likelihood of success. This parametrization allows for variation in question difficulty, and in particular, allows for both very hard questions, where even the top students have great difficulty, and very easy questions, where even the worst students have a high chance of selecting the correct answer. In addition, it maintains the effect of student ability: k is increasing in s, student ability, regardless of question difficulty.

A.3.2. *Estimation*

The model can be estimated through maximum likelihood. Let $x_{m,n} \in \{Correct, Incorrect, Skip\}$ denote the outcome of student m in question n; the data consists of question outcomes, $x_{m,n}$, for students $m = 1, ..., M$ and questions $n = 1, ..., N$. The probability of each outcome can be found, given $(k_{m,n}, c_m)$, or equivalently (s_m, q_n, c_m), where $k_{m,n}$ is the effective ability of student m in question n, and is equal to $\frac{q_n+s_m}{q_n}$, where q_n is the difficulty of the n^{th} question and s_m is the ability of the m^{th} student. The cutoff of the m^{th} student is c_m. We denote this probability as $P(x_{m,n}|s_m, q_n, c_m)$.[75] The estimates of student abilities $\{s_m\}_{m=1}^M$, the student risk aversion cutoffs $\{c_m\}_{m=1}^M$, and the question difficulties $\{q_n\}_{n=1}^N$ come

[73] With the penalty for incorrect answers set to 0.25, and five possible answers, we cannot distinguish between $c \in [0, 0.2]$, as the student will always answer every question. With alternative structures, the range of cutoffs we can consider will be different.

[74] The choice of scale parameter/support is without loss of generality is assumed to have identical support as previously.

[75] While Section 4 featured a correct answer and four incorrect answers, alternative numbers of incorrect answers can be considered. It is possible to have different numbers of possible answers in the same test; one may use a CARA utility function to find cutoffs for questions with different numbers of answers which feature the same level of risk aversion.

from the following:[76]

$$\max_{s_m, q_n, c_m} \sum_{n=1}^{N} \sum_{m=1}^{M} \log P(x_{m,n} | s_m, q_n, c_m). \tag{12}$$

A.3.3. *Data: Mock Exams*

The estimation procedure is applied to a mock exam held by a "dersane" (schools that prepare students for the exam). The exam consists of 120 questions: 30 for each subject, Turkish, social science, math and science. The same questions are shuffled to create two versions of the exam to prevent cheating.[77] We observe 30 students taking one version of the exam. We first apply the estimation procedure to the Turkish section of the exam, then to the social science section of the exam, then to both at the same time [78]

We recover question difficulties and student characteristics. Of particular importance is the relationship between scores and estimated abilities of students. A combination of the two abilities (Turkish and Social Science) gives us a measure of quality of the student. Figure A.11 depicts the correlation between ability ranking and score (with negative marking) as well as ability and the number of questions answered correctly.[79] As the figure shows, the three measures are highly correlated, but not perfectly so. In particular, the score ranking (one point for a correct answer, minus a quarter point for an incorrect answer) is closer to the 45 degree line than is the rank by total number of correct answers.[80] The reasons for this are discussed below.

With item response data, one can recover the ability parameters of students and question difficulties by using our model or the Rasch model. However, in addition to accommodating for skipping behavior, our model has some important features. In the Rasch model, the number of questions that a student answers correctly is a sufficient statistic for ability. For example, if a student gets 1 out of 2 questions correct, we can find ability, without knowing which questions. However, in our model, it is important which questions you answer correctly, and which questions you answer incorrectly (as well as skip). Suppose that one question is very easy, and one question is very challenging. Which student would likely have the higher ability - the one who answers the easy question correctly and not the difficult question, or the one who answers the difficult question correctly but not the easy question? Keeping in mind that a student with minimal ability can still have a one in five chance of correctly answering the question (random guess), we would say that the former student would have a higher estimated ability; the latter student could not answer the easy question correctly and simply got lucky with the hard question. Our estimation procedure incorporates this intuition.

Related to the previous observation, the fraction of students who get a question correct is not sufficient for describing difficulty. It is important to know which students get it right, and which students get it wrong. While the fraction getting it correct is highly correlated with difficulty, more so in large samples, it is not perfect. For small samples, there is valuable information contained in the identity of the students who correctly answer.

The Rasch model also does not distinguish between incorrect answers and a skip, while our model does. Suppose that there are twenty questions and two students with the same ability. However one is more risk averse than the other. As a result, the risk averse student skips questions and so has a lower fraction of correct questions on average. The Rasch model in this case would wrongly deem the risk averse student to have a lower ability. Risk preferences can, and do, vary across individuals. By incorporating skipping behavior, we can more accurately compare student's abilities. Our model, will account for differences in

[76]Note that it is necessary to make a normalization. We cannot identify the absolute difficulty of questions, only the relative difficulties. Without loss of generality, normalize the difficulty of the first question, $q_1 = 1$.

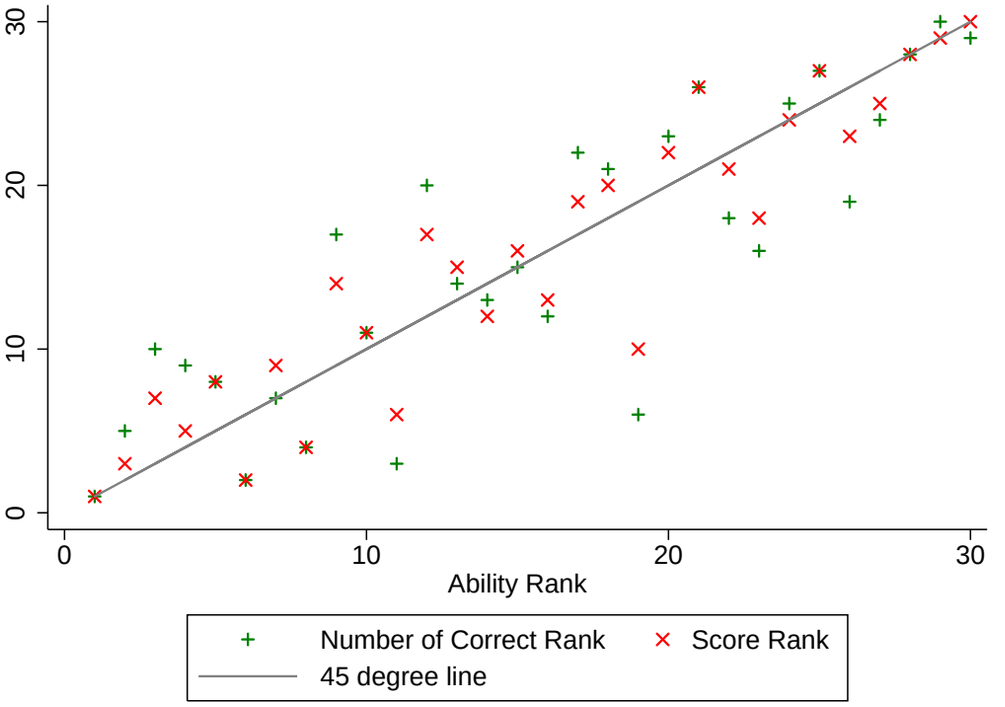
[77]We don't observe questions that correspond to each other in both versions. Therefore, we use one version of the exam.

[78]Here a student m is characterized by (s_m^T, s_m^{SS}, c_m). We normalize the first question of both exams to have a difficulty of 1 so that the difficulty of each question in each part of the exam is relative to the first question. Following this, the difficulties could be easily rescaled to be relative to the average difficulty for ease of interpretation.

[79]The number of questions answered correctly is important for the Rasch model.

[80]While the correlation between ability ranking and the score ranking is 0.929, the correlation between ability ranking and the number of correct ranking is 0.844.

Figure A.11: Ranking of students by estimated ability compared to ranking based on observables: number of correct answers and total score.



and the impact of risk preferences.

A.4. *Heterogenous Question Difficulty*

To investigate the impact of question difficulty heterogeneity on estimates of risk aversion, we simulated scores under an assumption of difficulty heterogeneity, and then applied the estimation approach set out in Section 4 to the simulated scores. We use the mock exam data that includes item by item responses of the students which is also used in subsection A.3. Firstly, we took the question difficulty estimates from the estimation of Turkish and Social Science sections of mock exam data, 30 questions each, as set out in section A.3. To generate 45 questions we appended questions equal to the average difficulty of the 1st and 2nd hardest questions, the 3rd and 4th hardest, and so on. The average difficulty of each section was normalized to one. We then simulated section scores for each of the students in the original OSS 1st time taker dataset, using parameter estimates for ability parameters and choosing selected cutoffs.

TABLE A.7

BIAS IN RISK AVERSION DUE TO QUESTION DIFFICULTY

	Female		Male	
Expected Score Interval	Actual	Estimated	Actual	Estimated
(0,90)	0.21	0.225	0.21	0.229
		(0.004)		(0.003)
[90,100)	0.22	0.236	0.22	0.236
		(0.001)		(0.001)
[100,110)	0.22	0.237	0.22	0.236
		(0.001)		(0.001)
[110,120)	0.23	0.244	0.23	0.244
		(0.001)		(0.001)
[120,130)	0.23	0.242	0.23	0.242
		(0.002)		(0.001)
[130,140)	0.24	0.251	0.24	0.249
		(0.003)		(0.003)
[140,inf)	0.24	0.242	0.24	0.245
		(0.007)		(0.006)

REFERENCES

AGNEW, J. R., L. R. ANDERSON, J. R. GERLACH, AND L. R. SZYKMAN (2008): "Who chooses annuities? An experimental investigation of the role of gender, framing, and defaults," *The American Economic Review*, 98(2), 418–422.

AHMADI, A., AND N. A. THOMPSON (2012): "Issues Affecting Item Response Theory Fit in Language Assessment: A Study of Differential Item Functioning in the Iranian National University Entrance Exam.," *Journal of Language Teaching & Research*, 3(3).

AKYOL, P., K. KRISHNA, AND S. LYCHAGIN (2022): "Deconstructing the Placement Gender Gap: Performance versus Preferences," .

AKYOL, P., K. KRISHNA, AND J. WANG (2021): "Taking PISA seriously: How accurate are low-stakes exams?," *Journal of Labor Research*, 42(2), 184–243.

BAKER, E. L., P. E. BARTON, L. DARLING-HAMMOND, E. HAERTEL, H. F. LADD, R. L. LINN, D. RAVITCH, R. ROTHSTEIN, R. J. SHAVELSON, AND L. A. SHEPARD (2010): *Problems with the use of student test scores to evaluate teachers*, vol. 278. Economic Policy Institute Washington, DC.

BALDIGA, K. (2014): "Gender Differences in Willingness to Guess," *Management Science*, 60(2), 434–448.

BECKER, W. E., AND C. JOHNSTON (1999): "The relationship between multiple choice and essay response questions in assessing economics understanding," *Economic Record*, 75(4), 348–357.

BELZIL, C., AND M. LEONARDI (2013): "Risk aversion and schooling decisions," *Annals of Economics and Statistics/Annales d'économie et de statistique*, pp. 35–70.

BEN-SHAKHAR, G., AND Y. SINAI (1991): "Gender Differences in Multiple-Choice Tests: The Role of Differential Guessing Tendencies," *The Journal of Educational Measurement*, 28(1), 23–35.

BEREBY-MEYER, Y., J. MEYER, AND O. M. FLASCHER (2002): "Prospect theory analysis of guessing in multiple choice tests," *Journal of Behavioral Decision Making*, 15(4), 313–327.

BERNARDO, J. (1998): "A Decision Analysis Approach to Multiple-Choice Examinations," *Applied Decision Analysis*, IV, 195–207.

BICKEL, J. E. (2010): "Scoring rules and decision analysis education," *Decision Analysis*, 7(4), 346–357.

BUDESCU, D., AND M. BAR-HILLEL (1993): "To guess or not to guess: A decision-theoretic view of formula scoring," *Journal of Educational Measurement*, 30(4), 277–291.

BURGOS, A. (2004): "Guessing and gambling," *Economics Bulletin*, 4(4), 1–10.

BURSZTYN, L., T. FUJIWARA, AND A. PALLAIS (2017): "'Acting Wife': Marriage Market Incentives and Labor Market Investments," *American Economic Review*, 107(11), 3288–3319.

BUSER, T., M. NIEDERLE, AND H. OOSTERBEEK (2014): "Gender, competitiveness and career choices," *The Quarterly Journal of Economics*, 129(3), 1409–1447.

CHARNESS, G., AND U. GNEEZY (2012): "Strong evidence for gender differences in risk taking," *Journal of Economic Behavior & Organization*, 83(1), 50–58.

COFFMAN, K. B., AND D. KLINOWSKI (2020): "The impact of penalties for wrong answers on the gender gap in test scores," *Proceedings of the National Academy of Sciences*, 117(16), 8794–8803.

CROSON, R., AND U. GNEEZY (2009): "Gender Differences in Preferences," *Journal of Economic Literature*, 47(2), 448–474.

DIRER, A. (2020): "Efficient scoring of multiple-choice tests," *Available at SSRN 3546770*.

DUFFIE, D., AND K. J. SINGLETON (1993): "Simulated Moments Estimation of Markov Models of Asset Prices," *Econometrica*, 61(4), pp. 929–952.

EBENSTEIN, A., V. LAVY, AND S. ROTH (2016): "The long-run economic consequences of high-stakes examinations: evidence from transitory variation in pollution," *American Economic Journal: Applied Economics*, 8(4), 36–65.

ECKEL, C. C., AND P. J. GROSSMAN (2008a): "Forecasting risk attitudes: An experimental study using actual and forecast gamble choices," *Journal of Economic Behavior & Organization*, 68(1), 1–17.

ECKEL, C. C., AND P. J. GROSSMAN (2008b): "Men, Women, and Risk Aversion: Experimental Evidence," *Handbook of Experimental Economics*, 1(113), 1061–1073.

ESPINOSA, M. P., AND J. GARDEAZABAL (2010): "Optimal correction for guessing in multiple-choice tests," *Journal of Mathematical Psychology*, 54(5), 415–425.

ESPINOSA, M. P., AND J. GARDEAZABAL (2013): "Do Students Behave Rationally in Multiple Choice Tests? Evidence from a Field Experiment," *Journal of Economics and Management*, 9(2), 107–135.

FREDERIKSEN, N. (1984): "The real test bias: Influences of testing on teaching and learning.," *American Psychologist*, 39(3), 193.

FUNK, P., AND H. PERRONE (2016): "Gender Differences in Academic Performance: The Role of Negative Marking in Multiple-Choice Exams," *CEPR Discussion Paper No. DP11716*.

GERBING, D. W., AND J. C. ANDERSON (1988): "An updated paradigm for scale development incorporating unidimensionality and its assessment," *Journal of marketing research*, 25(2), 186–192.

GOURIEROUX, C., AND A. MONFORT (1997): *Simulation-based econometric methods*. Oxford University Press.

GRISELDA, S. (2022): "The Gender Gap in Math: What are We Measuring?," *Available at SSRN 4022082*.

HANSEN, L. P. (1982): "Large sample properties of generalized method of moments estimators," *Econometrica: Journal of the Econometric Society*, pp. 1029–1054.

IRIBERRI, N., AND P. REY-BIEL (2021): "Brave boys and play-it-safe girls: Gender differences in willingness to guess in a large scale natural field experiment," *European Economic Review*, 131, 103603.

KAMAS, L., AND A. PRESTON (2012): "The importance of being confident; gender, career choice, and willingness to compete," *Journal of Economic Behavior & Organization*, 83(1), 82–97.

KARLE, H., D. ENGELMANN, AND M. PEITZ (2022): "Student performance and loss aversion," *The Scandinavian Journal of Economics*, 124(2), 420–456.

KUBINGER, K. D., S. HOLOCHER-ERTL, M. REIF, C. HOHENSINN, AND M. FREBORT (2010): "On Minimizing Guessing Effects on Multiple-Choice Items: Superiority of a two solutions and three distractors item format to a one solution and five distractors item format," *International Journal of Selection and Assessment*, 18(1), 111–115.

LYNN, R. (1992): "Sex differences on the differential aptitude test in British and American adolescents," *Educational Psychology*, 12(2), 101–102.

MANIAN, S., AND K. SHETH (2021): "Follow my lead: Assertive cheap talk and the gender gap," *Management Science*, 67(11), 6880–6896.

MILLER, J. B., AND A. SANJURJO (2018): "Surprised by the hot hand fallacy? A truth in the law of small numbers," *Econometrica*, 86(6), 2019–2047.

OECD (2009): *PISA Data Analysis Manual: SPSS, Second Edition*chap. The Rasch Model. OECD, Paris.

PEKKARINEN, T. (2015): "Gender differences in behaviour under competitive pressure: Evidence on omission patterns in university entrance examinations," *Journal of Economic Behavior & Organization*, 115(C), 94–110.

RASCH, G. (1993): *Probabilistic models for some intelligence and attainment tests.* ERIC.

RIENER, G., AND V. WAGNER (2017): "Shying Away from Demanding Tasks? Experimental Evidence on Gender Differences in Answering Multiple-Choice Questions," *Economics of Education Review*.

SAYGIN, P. O., AND A. ATWATER (2021): "Gender differences in leaving questions blank on high-stakes standardized tests," *Economics of Education Review*, 84, 102162.

VAN VELDHUIZEN, R. (2016): "Gender differences in tournament choices: Risk preferences, overconfidence or competitiveness?," *Journal of the European Economic Association*.